PERSON LEVEL ANALYSIS IN LATENT GROWTH CURVE MODELS

Ruth E. Baldasaro

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Psychology (Quantitative).

Chapel Hill
2012

Approved by:

Daniel J. Bauer

Patrick J. Curran

A. T. Panter

Andrea M. Hussong

David Thissen

ABSTRACT

RUTH E. BALDASARO: Person Level Analysis in Latent Growth Curve Models
(Under the direction of Daniel J. Bauer)

Latent growth curve modeling is an increasingly popular approach for evaluating longitudinal data. Researchers tend to focus on overall model fit information or component model fit information when evaluating a latent growth curve model (LGCM). However, there is also an interest in understanding a given individual's level and pattern of change over time, specifically an interest in identifying observations with aberrant patterns of change. Thus it is also important to examine model fit at the level of the individual. Currently there are several proposed approaches for evaluating person level fit information from a LGCM including factor score based approaches (Bollen & Curran, 2006; Coffman & Millsap, 2006) and person log-likelihood based approaches (Coffman & Millsap, 2006; McArdle, 1997). Even with multiple methods for evaluating person-level information, it is unusual for researchers to report any examination of the person level fit information. Researchers may be hesitant to use person level fit indices because there are very few studies that evaluate how effective these person level fit indices are at identifying aberrant observations, or what criteria to use with the indices. In order to better understand which approaches for evaluating person level information will perform best for LGCMs, this research uses simulation studies to examine the application of several person level fit indices to the detection of three types of aberrant observations including: extreme trajectory aberrance, extreme variability aberrance, and functional form aberrance. Results indicate that examining factor score estimates directly can help to identify extreme trajectory aberrance, while approaches examining factor score residuals or examining a person log-likelihood are better at identifying extreme variability aberrance. The performance of these approaches improved with more observation times and higher

communality. All of the factor score estimate approaches were able to identify functional form aberrance, as long as there were a sufficient number of observation times and either higher communality or a greater difference between the functional forms of interest.

Acknowledgements

This work could not have been done without the excellent training and mentorship provided by all of the faculty in the L.L. Thurstone Psychometic Lab. I have learned a tremendous amount from Dr. Robert MacCallum, Dr Patrick Curran, Dr. Abigail Panter, Dr. David Thissen, Dr. Sy-Minn Chow, and my advisor Dr. Daniel Bauer. I'm especially indebted to Dr. Daniel Bauer for being a patient teacher, who taught me how to focus my ideas and showed me that some of the best research ideas come from collaborating with other researchers.

In addition to the lab, I was also fortunate to work with Dr. Andrea Hussong, Dr. Michael Shanahan and Dr. Andrea Howard who all challenged me to be thoughtful about the purpose of my work and how to communicate it to others. I am also thankful for all the support and encouragement I received from the members of the Bauer Lab. I couldn't have made it through graduate school without watching Dr. Sonya Sterba and Dr. Nisha Gottfredson go through it first.

Lastly, this dissertation would not have been completed without the support of my family and my faith. I'm especially thankful for the continuous encouragement and frequent babysitting from the Baldasaros and for my son Calvin who has been a constant reminder of how hard work learning new things results in greater freedom. I'm also greatly indebted to my husband Nick, whose endless support kept me going when I wanted to give up. He made sure I finished this document. Finally, I'm thankful to God who has blessed me with many gifts and wonderful people to help me learn to use them.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figure

Chapter 1

Introduction

Latent growth curve modeling is an increasingly popular approach for evaluating longitudinal data because it provides information about how individuals change over time both on average, as well as how much variability there is in change over time. It also provides a method which allows researchers to test whether covariates can explain why some individuals appear to have different levels or patterns of change over time. Researchers tend to focus on overall model fit information or component model fit information when evaluating a latent growth curve model (LGCM). However, given the interest in understanding a given individual's level and pattern of change over time, it is also important to examine model fit at the level of the individual. Examining individual-level information can quantify how well or poorly a given model explains an individual's observations, identify outliers or subgroups of individuals who are not well represented by a given model, and/or provide evidence of model misspecification. Thus, examining person-level data can help answer the question: Are there individuals in a set of data for whom a given model does not represent observations well?

This is an important question to answer because the goal of fitting a statistical model like a LGCM is to test a theory for how a set of variables is related. There are many ways that an individual's observations may appear to be poorly represented by a statistical model. This typically occurs when the model implies a relationship among the observed variables that is systematically different from the observed data. One example of an individual whose observations are systematically different from what a model implies about the individual would be someone whose observed depression scores are consistently much higher than average in a college student sample. This individual could represent normal variability in that population or the individual may represent an observation from a clinically depressed population

that is distinct from the population of college students that the researcher intended to sample. Another way that an individual's observations may appear systematically different is when the model-implied functional form does not represent the individual's observed data well. For example, the model may imply a positive linear trend in reading ability over time, but the data for an individual shows a nonlinear relationship. This difference between the model-implied functional form and the observed pattern in the data could occur simply by chance, if the individual's reading ability happened to improve slowly initially and then accelerated over time, or it may be evidence that the model is misspecified and that the theory the model is testing is incorrect about how reading ability changes over time. These are just a few examples of how a model may not represent an individual's observations well. Whenever a researcher finds these individuals in his or her data, the researcher should contemplate why these individuals are present in the data. Some possibilities are (1) the individuals represent observations from different populations, (2) there is a characteristic of the individuals that can explain their differences, or (3) the model has been misspecified for these individuals. Alternatively, it could be that these observations appear to be aberrant simply by chance. Thus, having aberrant observations may provide information on different populations in a given sample, suggest a need to bring covariates into a model, or suggest a need to change how a model is specified. As a result, it is important to have methods to identify these aberrant individuals.

Currently there are several proposed approaches for evaluating person level information from a LGCM including factor score based approaches (Bollen & Curran, 2006; Coffman & Millsap, 2006) and person log-likelihood based approaches (Coffman & Millsap, 2006; McArdle, 1997). Even with multiple proposed methods for evaluating person-level information, it is unusual for researchers to report any examination of the person-level information despite the recommendation to do so in several text books and articles (Bollen & Curran, 2006; Carrig, Wirth, & Curran, 2004; Coffman & Millsap, 2006; McArdle, 1997; Preacher, et al, 2008). The lack of reported examination of person-level information in applications of LGCMs is surprising, given the extensive coverage of person-level diagnostics in multilevel models

(Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999), which are similar models to LGCMs (Bauer, 2003; Chou, Bentler, & Pentz, 1998; Curran, 2003; MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997). Also, the increasing interest in examining heterogeneity in longitudinal data (Nagin, 1999, 2005; Muthén, & Muthén, 2000; Muthén, 2001, 2002) suggests that researchers would be interested in procedures to examine model results at the level of the individual.

Because of the recommendations to perform person level analysis and the interest in understanding model fit at the level of the individual, it is unclear why researchers do not report person level analysis in applications of LGCMs. A potential reason for the absence of person level analysis in LGCMs could be the lack of information regarding the performance of the approaches for person level analysis. There are many strategies for evaluating person level information, but there are very few studies that evaluate how well the approaches perform in practice, especially for LGCMs. There is also very little information on how these approaches perform relative to each other. It would be useful to know which approaches are better for identifying different types of aberrant individuals. To better understand which approaches for evaluating person level information will perform best for LGCMs, I will first provide an introduction to latent growth curve models. Next, I will describe in greater detail the types of aberrant observations that could appear in longitudinal data. Then I will review the proposed approaches for evaluating person level information in order to identify aberrant observations from a variety of statistical models, as well as information about the performance of these approaches. After reviewing the proposed approaches, I will identify the information that is needed to understand how to apply these approaches to a LGCM effectively. Last, I will provide an overview and hypotheses for the current research, which aims to provide evidence for how well these approaches work, and under what conditions they perform better or worse.

Introduction to Latent Growth Curve Modeling

Latent growth curve modeling is one approach for modeling changes in a variable over time. In a LGCM, repeated observations of a variable (or variables) are assumed to be the result of a systematic

underlying trajectory of change over time with the addition of random time-specific error. This model can be represented as

$$\boldsymbol{y}_i = \boldsymbol{\Lambda}\boldsymbol{\eta}_i + \boldsymbol{\varepsilon}_i \quad (1)$$

where $\boldsymbol{y}_i$ represents a vector of the observed repeated measurements for person i, $\boldsymbol{\Lambda}$ is the matrix of weights or factor loadings representing the functional form of the latent trajectory, $\boldsymbol{\eta}_i$ is a vector of the latent trajectory parameters for person i, and $\boldsymbol{\varepsilon}_i$ is a vector of time-specific error for person i. The vector of latent trajectory parameters ($\boldsymbol{\eta}_i$) can be represented as

$$\boldsymbol{\eta}_i = \boldsymbol{\mu}_\eta + \boldsymbol{\zeta}_i \quad (2)$$

where $\boldsymbol{\mu}_\eta$ represents a vector of mean trajectory parameters, and $\boldsymbol{\zeta}_i$ represents a vector containing the deviations from the mean trajectory parameters for person i. It should be noted that Equations 1 and 2 are for an unconditional LGCM. For a conditional LGCM, predictors of the observed data could be included in Equation 1, and predictors of the individual latent trajectory parameters could be included in Equation 2.

The mean and covariance structure derived from Equation 1 and 2 can be used to obtain parameter estimates for the LGCM. The covariance structure is

$$\boldsymbol{\Sigma}_{yy} = \boldsymbol{\Lambda}\boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}_\varepsilon \quad (3)$$

where $\boldsymbol{\Sigma}_{\eta\eta}$ is the covariance matrix for the latent variables, $\boldsymbol{\Sigma}_{yy}$ is the covariance matrix for the observed variables, and $\boldsymbol{\Theta}_\varepsilon$ is the covariance matrix for the disturbances or error variances. The mean structure is

$$\boldsymbol{\mu}_y = \boldsymbol{\Lambda}\boldsymbol{\mu}_\eta \quad (4)$$

Together, Equations 3 and 4 represent model-implied mean and covariance structure of the repeated measures. Keeping this model for longitudinal data in mind, I will now describe several types of

aberrant observations that can appear in longitudinal data and some potential explanations for why aberrant observations occur.

Types of Aberrant Observations in Longitudinal Data

There are many ways that individuals can appear aberrant in longitudinal data. To help illustrate some of the more commonly observed types, I will first provide a description of the general pattern for the aberrant observations and then provide a few examples of why an individual may have a given type of aberrance. After each type of aberrant individual is described, I will provide a summary of some potential explanations for why a given individual may appear aberrant relative to the rest of the data.

One type of aberrant observation in longitudinal data is that of individuals who have extremely high or low levels relative to the average level in a given sample. Earlier I described an example of an individual with consistently high depression ratings. This type of aberrance could also appear in educational data as individuals who show either extremely high or extremely low academic ability relative to their peers. As with the depression data, these observations could simply occur as legitimate observations that appear at the extremes of the distribution of ability, or they could represent individuals who are from a fundamentally different population. As an example, certain individuals may appear to have extremely low reading ability because they are being tested in a language in which they are not native, or because of some developmental disability.

In addition to having an extremely high or low level relative to the average level, individuals could appear to be aberrant in longitudinal data because their rate of change over time is much faster or slower than average. For example, students who start out extremely low in academic performance but show dramatic improvement over time may appear to have a positive growth in ability over time that occurs much faster than average. As with the previous examples, this individual could represent the normal variability in academic performance, albeit extreme relative to his or her peers. Alternatively, this pattern of change could result from changes in motivation over time. For example, a student with higher

ability could be bored at younger ages and not take testing very seriously, resulting in choosing answers that cause the student to appear to have a lower than average ability. As the student gets older, the student may become more motivated, perhaps by the opportunities that become available to students who perform well on the test (e.g. more or better college opportunities), resulting in choosing answers that better reflect the student's true ability, which is higher than average. This would result in a pattern of observations with an extremely high positive change over time, which would appear aberrant relative to the other individuals in the sample, but which could be explained if motivation were measured and included in the model.

The previous two examples describe individuals whose longitudinal data appear to have extreme level or slope differences from the average level and slope observed in the sample. It is also possible to have some combination of these extremes. Overall, these types of aberrant observations could be described as extreme trajectory aberrance.

In contrast to extreme trajectory aberrance, it is also possible for an individual to have an overall trajectory that is close to the average trajectory observed in the sample, but appear aberrant because the observations at each time point show significantly higher variability relative to other individuals. For example, an individual may appear on average to have normal, stable depression scores over time based on his or her model implied trajectory. That individual could have observed depression scores that are at relatively normal levels at each time point or the individual could have observed depression scores that fluctuate over time between healthy low levels and clinically problematic high levels over time. In the second case, the individual would appear to be aberrant due to the high variability in their scores relative to other individuals in the sample. This variability could occur simply by chance, perhaps because the study happens to measure the individual at times when they are experiencing healthy low levels of depression, and then also shortly after a family member passes away or the individual loses his or her job. Alternatively, the individual could represent someone from a different population of individuals, perhaps

someone who has bipolar disorder. This type of aberrant observations could be described as extreme variability aberrance.

Another way an individual's observations may appear aberrant in longitudinal data is when the shape of the trajectory specified by the model does not represent the shape of the trajectory of the observed data for an individual well. Earlier I provided an example of an individual whose reading ability improved slowly initially and then accelerated over time, while the model specified a linear change in reading ability over time. A similar situation could occur when observing the progression of mental health over time. A study following individuals during and after treatment could test the hypothesis that improvement in mental health is linear over a short time period, while an individual's observations may show a nonlinear change over time. It could be that the nonlinearity in the observed data just occurred by chance, or it could be evidence that the pattern of change implied by the theory being tested is incorrect. This type of aberrance could be described as functional form aberrance.

Yet another possible form of aberrance in longitudinal data would be an individual who has one or a few observations that are extremely different from the rest of his or her observations. This could occur in an academic setting because the individual is very sick the day of testing or if the individual is only able to cheat on one or a few of the tests. In substance use research, this could occur if one or a few of the time points happen to occur during a period of time when the individual is experimenting with drugs, but for an individual who does not typically use substances. This type of aberrant observation could be described as aberrance due to time-specific outliers.

These different types of aberrant observations can cause problems when trying to understand the overall trends in longitudinal data. Whether or not an aberrant observation is problematic depends on why the aberrant observation appears in the data set. There are many possible explanations for why an individual's observations may not be well represented by a latent growth curve model. I will describe four potential explanations. The first explanation for why an individual's observations may appear aberrant is

because the individual is from a different population. For these individuals, a researcher may want to exclude them from the sample because the individuals are not from the population of interest. A second explanation for why an individual may appear aberrant is because there is some characteristic of these individuals that can explain the difference between what the model implies and the observed data. For these individuals, researchers would not want to exclude them from the sample, but covariates should be included in the model to explain why these individuals appear to be aberrant relative to the other individuals in the sample. A third explanation for why an individual may appear to be aberrant is because the model is incorrectly specified. In this situation, researchers may decide that the individuals provide evidence that a theory does not result in an accurate representation of the data, and therefore the theory needs to be revised. Finally, an individual may appear to be aberrant simply by chance. For these individuals the model is not misspecified, there is no characteristic of these individuals to explain their aberrance, and the individual is from the population of interest. In this situation, researchers would want to include the individuals because they are valid observations of how a construct is changing over time.

These potential patterns of aberrant observations in longitudinal data and potential explanations for aberrance are not intended to include all possible types or explanations of aberrance. They are intended to provide some examples of how and why aberrant observations may appear. Given that these patterns of observations may result in researchers making decisions about including or excluding individuals or covariates, or revising theory, it is important to be able to identify these aberrant individuals. The following sections will examine potential approaches that use person level analysis to identify aberrant individuals in LGCMs.

Current Approaches for Person Level Analysis in a LGCM

Based on a review of the previous work on person level analysis in LGCMs and the analyses done for other models, there appear to be two general approaches to person level analysis, factor score analysis and person level fit analysis. To better understand what each of these approaches can provide regarding person fit, I will describe the motivation for each approach and summarize some strategies for

implementing each approach. After presenting the approaches and strategies of implementing them, I will describe what research is needed in order to support the use of these strategies for identifying different types of aberrant observations in LGCMs.

Factor Score Analysis

There are several approaches for evaluating person level information using factor score estimates. These approaches either focus solely on the factor score estimates, or use the factor score estimates to generate predicted observations and calculate residuals for analysis. The motivation behind examining the factor score estimates directly is typically either to identify unusual individuals or subgroups at the level of the latent variables, or to check for any unusual patterns in the factor score estimates suggesting model misspecification. The motivation behind calculating predicted observations and residuals is typically to quantify the fit of the model in terms of how far the predicted observations are from the observed data. These quantifications of model fit provide an index of model fit at the level of the individual, which could potentially be used to compare models. To better understand how these factor score based approaches are implemented, I will first describe approaches for calculating factor score estimates, predicted values, and residuals, then describe how several researchers have proposed using these calculations to perform person level analysis.

There are several methods of calculating factor score estimates. In order to select a method for calculating factor score estimates one must understand both the general and specific properties of factor score estimates. The most important general property of factor score estimates is that they are predictions, not estimates, of latent variables. Factor score estimates are predictions because of factor indeterminacy. Factor indeterminacy occurs because there are more unknown variables than known variables in latent variable models. Factor indeterminacy results in the ability to find infinite sets of factor score estimates, factor loadings, and unique factors that can explain the pattern of relationships among observed variables equally well. This means that factor score estimates cannot be uniquely determined, although the degree of determinacy can vary (Guttman, 1955; McDonald, 1974). Because of factor indeterminacy, researchers

have been warned not to treat these scores as the true latent variable scores (Gorsuch, 1983), and not to make fine comparisons using factor score estimates (Bollen, 1989). However, Gorsuch (1983) argues that,

> "Indeterminacy of factor scores is only a serious problem if one reifies the factors from a given study into ultimate realities. The present approach is to consider factors as constructs that will, hopefully, aid in theory development. Constructs are always abstractions from data and never observed. The best one can expect with any construct is that a psychometrically good operational representative will be found for it (Nunnally, 1978)."

Thus, Gorsuch argues that as long as factor score estimates are a good representation of the latent factors, they can be used as predictions of the latent factors. Similarly, Raykov and Penev (2002) argue that the issue of factor indeterminacy, or lack of unique factor score estimates, does not necessarily imply that the information contained in a given set of factor score estimates lacks meaning. In addition to these arguments, there is empirical evidence that factor score estimates can be useful for identifying unusual observations or model misspecifications (Bollen & Arminger, 1991; McDonald & Bolt 1998; Raykov & Penev, 2002; Bauer, Baldasaro, & Gottfredson, 2012).

As stated previously, there are many ways to calculate factor score estimates (Grice, 2001a; 2001b). This study will focus on two approaches, the regression method, also known as the empirical Bayes method (Thomson, 1936,1951; Thurstone, 1935), and Bartlett's method, also known as the generalized least squares (GLS) method (Bartlett, 1937). I will now describe the specific properties of these two methods.

Factor score estimates regression method. Factor score estimates can be generated by the regression method (Thomson, 1936, 1951; Thurstone, 1935). Regression factor score estimates are calculated as

$$\widehat{\boldsymbol{\eta}}_{rm_i} = \boldsymbol{\Sigma}_{\eta\eta}\boldsymbol{\Lambda}'\widehat{\boldsymbol{\Sigma}}_{yy}^{-1}\left[\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_y\right] + \boldsymbol{\mu}_\eta \qquad (5)$$

where $\boldsymbol{\Sigma}_{\eta\eta}$ is the covariance matrix for the latent variables, $\boldsymbol{\Lambda}$ is the factor loading matrix, $\widehat{\boldsymbol{\Sigma}}_{yy}$ is the model-implied covariance matrix for the observed variables, $\boldsymbol{y}_i$ is a vector of observed variables for

person i, $\widehat{\boldsymbol{\mu}}_y$ is a vector of the model-implied means for the observed variables and $\boldsymbol{\mu}_\eta$ is a vector of the latent variable means. The motivation behind the regression method was to develop a way to produce an equation that would calculate the best prediction of the factor score for any person (Bartholomew, Deary, & Lawn, 2009). These are equivalent to empirical Bayes estimates (Bartholomew & Knott, 1999; Skrondal & Rabe-Hesketh, 2004). These predictions are shrunken towards the distribution of the latent variables in the fitted model. The level of shrinkage depends on factor determination, which depends on the informativeness of the observed data. As the number and communality of measured variables increases factor determination increases (Grice, 2001a). In the multilevel modeling (MLM) literature, the shrinkage of empirical Bayes estimates has been found to be greater for individuals with fewer observations, and when within-subject variability is large relative to between-subjects variability (Verbeke & Molenburghs, 2000). For longitudinal data, fewer observations would be equivalent to fewer time points, and greater within-subject variability relative to between-subject variability, would be equivalent to greater time-specific variability relative to latent trajectory variability. Empirical Bayes estimates are often described as 'shrunken' because they tend to shrink towards the mean, which makes them both more biased but also more precise (Hox, 2010). Gorsuch (1983) notes that these factor score estimates have the highest correlation with actual factors.

Factor score estimates Bartlett's method. Factor score estimates can also be generated by Bartlett's method (Bartlett, 1937). Bartlett's factor score estimates are calculated as

$$\widehat{\boldsymbol{\eta}}_{GLS_i} = (\boldsymbol{\Lambda}'\boldsymbol{\Theta}_\varepsilon^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Theta}_\varepsilon^{-1}\left[\boldsymbol{y}_i - \widehat{\boldsymbol{\mu}}_y\right] + \boldsymbol{\mu}_\eta \qquad (6)$$

This method is sometimes referred to as the generalized least squares (GLS) method (Bollen & Curran, 2006) or the maximum likelihood method (Skrondal & Rabe-Hesketh, 2004). Bartlett's scores minimizes the size of the unique factors by minimizing the sum of the squared errors divided by their standard deviations. In contrast to the regression method, the motivation behind Bartlett's method was to calculate a factor score estimate for a specific person with a given set of observed scores, with the properties that

the prediction be unbiased and have minimized variance (Bartholomew, Deary, & Lawn, 2009). By minimizing the unique variances, Bartlett's approach results in factor score estimates that are conditionally unbiased (Goursuch, 1983; Skrondal & Rabe-Hesketh, 2004). This means that unlike the regression method, which is known to provide biased predictions, Bartlett's approach should provide unbiased predictions given that the model is properly specified.[1]

Bartlett's method optimizes a different set of criteria than the regression method, which means that applying the two approaches to the same model with the same data will not result in identical factor score estimates for a given person. The rank order of individuals in terms of their factor score estimates will differ across these factor score methods unless the individuals in the data are independent and the latent factors are uncorrelated (McDonald, 2011). Researchers are typically encouraged to use whichever approach is appropriate given the aims of their factor score analysis (Tucker, 1971). McDonald (2011) has argued that it is not clear which of these factor score approaches should be used when examining longitudinal models such as the LGCM. Because these approaches optimize different criteria, it is likely that there are some person level analysis approaches that would favor using Bartlett's factor score estimates over regression factor score estimates, while other person level analysis approaches may favor regression factor score estimates over Bartlett's. Thus, in order to determine with factor score approach is appropriate for a given person level analysis approach, both of these approaches for calculating factor score estimates need to be examined.

Residuals from factor score estimates. In addition to calculating factor score estimates, researchers can use the scores to generate predicted values and residuals for the researcher's observed data. Predicted values for a LGCM can be calculated as

[1] Note that Equation 5 and 6 would represent the factor score equations for unconditional LGCMs. These equations could be extended to conditional LGCMs with exogenous predictors by replacing $\boldsymbol{\mu}_{\eta}$ with a vector of conditional means. Bringing in relevant exogenous predictors may reduce the bias of the regression method (McDonald, 2011).

$$\hat{\boldsymbol{y}}_i = \boldsymbol{\Lambda}\,\hat{\boldsymbol{\eta}}_i \qquad (7)$$

where $\hat{\boldsymbol{y}}_i$ is a vector of the predicted values, and $\hat{\boldsymbol{\eta}}_i$ is the vector of factor score estimates for person i. The factor score estimates can be calculated using the factor score methods described in Equation 5 and 6, or from any other factor score method. The predicted values from different factor score methods will not be equal.

Residuals are calculated using the following equation

$$\hat{\boldsymbol{\varepsilon}}_i = \boldsymbol{y}_i - \hat{\boldsymbol{y}}_i \qquad (8)$$

where $\hat{\boldsymbol{\varepsilon}}_i$ is a vector of residuals for person i. This residual calculation can be used for any factor score method. As with the predicted values, the residuals for person i will not be identical for different factor scoring methods.

Person level analysis using factor score estimates. Many approaches have been proposed to use factor score estimates, predicted values, or residuals for person level analysis. I will first describe the approaches proposed in the LGCM literature. In addition to the LGCM literature, I will also highlight some approaches described in the multilevel modeling and structural equation modeling literature that may be useful in LGCM applications.

In their book on latent growth curve modeling, Bollen and Curran (2006) provide instructions for how to evaluate person level diagnostics using factor score estimates. Bollen and Curran suggest that researchers use factor score estimates of the latent intercepts and slope factors to perform diagnostics in the same way as how ordinary least squares (OLS) estimates of the intercept and slope for each individual are used. They give an example of how to perform a person level analysis with OLS estimates of the latent intercept and slope factors, and also provide equations for several popular factor scoring methods. In their example of OLS diagnostics, they examine plots of the estimated intercepts and slopes for extreme estimates. They also use the predictions from factor score estimates to calculate R-squared values

as a means of quantifying the closeness of a linear model to the observed data. They argue that a strong linear trend would be indicated by R-squared values of 0.7 or more. Thus, Bollen and Curran propose methods for using factor score estimates both to identify unusual observations and to quantify model fit. However, they did not provide any information on the limitations of these approaches, or how well these approaches would work.

Coffman and Millsap (2006) also describe using factor score methods to produce estimates of individual latent trajectories. First, they discuss using factor score estimates to calculate predicted latent variables and then generate predicted observations which can then be used in a residual analysis. For an example of how to implement a residual analysis they cite the work done by Bollen and Arminger (1991) on person-specific residual analysis in structural equation modeling (SEM). Coffman and Millsap propose that Bollen and Arminger's residual analysis for SEM could also be used for person level diagnostics for LGCMs. However, Coffman and Millsap did not implement this residual analysis approach with a LGCM or evaluate its performance.

In addition to describing a residual analysis approach, Coffman and Millsap (2006) also discuss the procedure described by Meredith and Tisak (1990) for obtaining empirical Bayes estimates for the latent variables, and how these estimates could be used to evaluate person level diagnostics. Specifically, Coffman and Millsap describe how the empirical Bayes estimates could be used to calculate predicted latent trajectories for each individual, and then show how the predicted latent trajectories could be compared to the observed data by computing a root mean squared residual for each individual (RMSR_i)

$$\mathrm{RMSR}_i = \sqrt{\frac{\hat{\boldsymbol{\varepsilon}}_i' \hat{\boldsymbol{\varepsilon}}_i}{\mathrm{T}_i}} \tag{9}$$

where T_i is the total number of time points for person i. The RMSR_i is a numerical index of how close the model implied trajectory is to the observed longitudinal data for a given person. As with the residual

analysis, this proposed RMSR_i approach to person level diagnostics was not implemented or evaluated for effectiveness, nor has it been compared to other proposed strategies.

In addition to the LGCM literature, the literature on multilevel models (MLMs) can offer some strategies for using factor score estimates to evaluate LGCMs at the level of the individual. Several authors discuss the similarities and differences between MLMs and LGCMs (Bauer, 2003; Chou, Bentler, & Pentz, 1998; Curran, 2003; MacCallum, Kim, Malarkey, & Kiecolt-Glaser, 1997). A multilevel model and a LGCM applied to the same data can result in identical parameter estimates when the models have the same constraints. Given this similarity, it is worthwhile to examine what analysis of the person level is recommended in multilevel modeling. In a longitudinal MLM, examining person level information involves estimating the level 2 random effects or residuals. These level 2 estimates or residuals can be examined for outliers, subgroups, or assumption violations. Estimates of the person level random effects are usually calculated using a Bayesian approach and the estimates are referred to as "empirical Bayes" estimates (Langford & Lewis, 1998; Snijders & Bosker, 1999; Verbeke & Molenberghs, 2000; Hox, 2010). However, some authors also discuss using ordinary least squares estimates (Raudenbush & Bryk, 2002; Skrondal & Rabe-Hesketh, 2004). Once person level estimates are obtained, a variety of plots can be used to examine the distributions of the estimates or their residuals (e.g. histograms, boxplots, and scatterplots) for outliers or subgroups (Langford & Lewis, 1998; Snijders & Bosker, 1999; Verbeke & Molenberghs, 2000; Raudenbush & Bryk, 2002; Hox, 2010). The normality of the estimates can be evaluated by examining skew or kurtosis, and residuals can be examined for normality using a normal Q-Q plot (Langford & Lewis, 1998). Some textbooks also recommend examining plots of unstandardized residuals and their relationship with relevant covariates to check for functional form, homoscedasticity, and influential observations (Snijders & Bosker, 1999; Raudenbush & Bryk, 2002). Similar to linear regression, there are formulas available for calculating Studentized residuals, leverage, and influence for a MLM (Langford & Lewis, 1998). These statistics can be compared to cut-off values to identify individuals or subgroups that are outliers or influential. These MLM person level diagnostics provide

some additional ideas for how to use factor score estimates for person level diagnostics in applications of LGCMs.

In contrast to the multilevel modeling research, structural equation modeling (SEM) research has not spent as much time on the issue of person level analysis. While MLM textbooks routinely discuss person level analysis, textbooks on SEM typically discuss evaluating models using overall model fit, component fit, or proportion of variance explained (Bollen, 1989; Bollen & Long, 1991; Kaplan, 2000; Schumacker & Lomax, 2004; Mulaik, 2009). Person level analysis is generally discussed in research examining how to use factor score estimates or factor score based residuals, to identify outliers or violations of model assumptions (Bollen & Arminger, 1991; McDonald & Bolt, 1998; Raykov & Penev, 2002), or how much a person contributes to model misfit using person log-likelihoods (Reise & Widaman, 1999; Sideris, 2006). Both the residual analysis and person log-likelihood approaches attempt to identify individuals who do not fit with a given model. I will discuss the person log-likelihood approaches later. In the subsequent sections I will highlight the SEM work on using factor score estimates for person level analysis.

The use of factor score estimates for person level analysis was demonstrated by Bollen and Arminger (1991) when they developed person-specific residual analysis in SEM. They proposed two factor score methods for calculating unstandardized residuals. The first method estimated unstandardized residuals with a weight matrix chosen so as to minimize the sum of squared residuals. This method is based on the regression method of calculating factor score estimates (Thomson, 1936, 1951; Thurstone, 1935). The second method estimated unstandardized residuals with a weight matrix chosen so as to minimize the sum of squared errors divided by their standard deviation. This method is based on the Bartlett's method of calculating factor score estimates (Bartlett, 1937). Bollen and Arminger (1991) also proposed ways to calculate variances and standardized residuals for each method. The variance for the residuals from the regression method is

$$\mathrm{VAR}(\hat{\boldsymbol{\varepsilon}}_{rm_i}) = \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\Sigma}}_{yy}^{-1}\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}} \qquad (10)$$

and for residuals from Bartlett's method the variance is

$$\mathrm{VAR}\left(\hat{\boldsymbol{\varepsilon}}_{GLS_i}\right) = \boldsymbol{\Theta}_{\boldsymbol{\varepsilon}} - \boldsymbol{\Lambda}(\boldsymbol{\Lambda}'\boldsymbol{\Theta}_{\boldsymbol{\varepsilon}}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}' \quad (11)$$

To standardized the residuals Bollen and Arminger used the following calculation

$$\hat{\varepsilon}_{std_{ji}} = \hat{\varepsilon}_{ji}/\sqrt{[\mathrm{VAR}(\hat{\boldsymbol{\varepsilon}}_i)]_{jj}} \quad (12)$$

where $\hat{\varepsilon}_{std_{ji}}$ is the *j*th residual of $\hat{\boldsymbol{\varepsilon}}_i$, which is the vector of residuals for person i, and $[\mathrm{VAR}(\hat{\boldsymbol{\varepsilon}}_i)]_{jj}$ is the *j*th diagonal element of $\mathrm{VAR}(\hat{\boldsymbol{\varepsilon}}_i)$. The same calculation can be used for the residual and variance estimates from either the regression or Bartlett's method. Bollen and Arminger provided recommendations for how to examine residuals graphically and with statistical criteria. Specifically, they recommended using a simultaneous test of statistical significance for the sample residuals for a person. They calculated the test statistic as follows

$$h_i = \hat{\boldsymbol{\varepsilon}}_i'[\mathrm{VAR}(\hat{\boldsymbol{\varepsilon}}_i)]^{-1}\hat{\boldsymbol{\varepsilon}}_i \quad (13)$$

where the residual ($\hat{\boldsymbol{\varepsilon}}_i$) and variance ($\mathrm{VAR}(\hat{\boldsymbol{\varepsilon}}_i)$) estimates can come from either the regression or Bartlett's method. The test statistic h_i follows a chi-distribution with degrees of freedom equal to the number of observed variables; for the latent curve model this will be the number of observed time points. Because h_i is calculated for each individual, Bollen and Arminger recommended that researchers apply a correction for multiple significance testing. They proposed several corrections, including a Bonferroni correction or simply selecting a different alpha level and dividing it by the number of significance tests examined.

Bollen and Arminger used simulations to examine how well the two approaches performed at identifying outliers. The simulation studies found that these approaches to estimating residuals were helpful at identifying unusual observations even at small sample sizes. It was emphasized that the methods proposed were one way to identify outliers, in this case, observations that were not well-predicted by the model. It is possible that some of the outliers identified with these methods would also be

influential observations. This would mean that when the observation is removed from the model, substantial changes in model parameters or overall fit would occur. The authors warn that examining residuals does not guarantee that all outliers or influential observations will be identified and suggest that there is a need to develop statistics similar to the regression diagnostics: DFBETAS, Cook's *D*, and DFITS. They also emphasized that as with any new procedure, more research would be needed to identify the strengths and limitations of their proposed procedures.

Similar to Bollen and Arminger (1991), McDonald and Bolt (1998) showed how to use the regression method first to obtain factor score estimates for calculating predicted observations and then to calculate residuals. McDonald and Bolt focus on how factor indeterminacy influences residual analysis in SEM. Specifically, they examined the detection of departures from the assumptions of linearity using residuals calculated by taking the difference between the predicted observations based on factor score estimates and the observed measures. They found that the factor indeterminacy did not hinder the ability to use residual analysis for this purpose. The authors argued that their findings provide evidence that residual analyses in SEM can be used to detect outliers or assumption violations even with factor indeterminacy.

Raykov and Penev (2002) were the first to specifically emphasize using factor score estimates to evaluate person level model fit. They created person level residuals by generalizing Bartlett's based residuals proposed by Bollen and Arminger (1991). They showed how the person level residuals can be used to evaluate model fit by examining plots of the residuals for trends that suggest model misspecification. Raykov and Penev (2002) defined person fit as the difference between the observed data and the predicted data based on the model estimates, and contrasted person fit to the traditional overall model fit indices that examine the difference between the observed and model implied covariance estimates.

In addition to the previously described work on residual analysis, several researchers have examined ways to calculate residual-based Mahalanobis distance (Ke-Hai, Wing Kam, & Reise, 2004), and case influence measures (Lee & Wang, 1996; Pek & MacCallum, 2011; Tanaka, Watadani, & Moon, 1991) for structural equation models. Pek and MacCallum (2011) provide a comprehensive overview of how to calculate several outlier indices common to linear regression in an SEM setting. One challenge to using the case influence measures is that they are typically computationally burdensome with large sample sizes, because they involve fitting the model of interest to the full sample and then again, removing one observation at a time. This burden may be remedied by incorporating these calculations into SEM software programs. Until then, researchers may prefer person fit indices that only require minimal computational burden, like the previously described factor score approaches or the person log-likelihood-based indices, which are described next.

Person Level Fit Analysis

In addition to factor score approaches, several researchers have discussed using a measure of person level fit to evaluate how well an individual's observations fit with a given model. Preacher, Wichman, MacCallum, and Briggs (2008, p. 21) recommend that researchers examine individual fit criteria in addition to global fit criteria. They argue that person fit should be examined in order to determine whether the individual growth curves are approximated well. Several articles propose using person log-likelihood indices to evaluate person level fit in a LGCM (McArdle & Bell, 2000; McArdle & Hamagami, 2001; Mehta & West, 2000). Only a few articles show how to implement an analysis using person log-likelihoods (Blozis & Cho, 2008; Coffman & Millsap, 2006; McArdle, 1997). Specifically, McArdle (1997) partitioned the overall log-likelihood into person log-likelihoods using a formula for a person log-likelihood described by Lange, Westlake and Spence (1976)

$$PLL_i = -\frac{1}{2}\left[p\ln(2\pi) + \ln\left|\hat{\Sigma}_{yy}\right| + \left(y_i - \hat{\mu}_y\right)'\hat{\Sigma}_{yy}^{-1}\left(y_i - \hat{\mu}_y\right)\right] \qquad (14)$$

Where p is the number of measured variables. The overall $-2LL$ for a given model, can be calculated by taking two times PLL_i and then taking the sum of the -$2PLL_i$ across the individuals in the sample. McArdle examined a plot of the person log-likelihoods to look for individuals who made greater contributions to the overall model misfit. After identifying individuals with large person log-likelihoods (relative to the other individuals), he then calculated the percent of the overall log-likelihood that was contributed by the unusual individuals. Thus, McArdle (1997) identified two ways for examining individual log-likelihoods for unusual observations: using a plot of the person log-likelihoods and calculating the percent of the overall log-likelihood a given individual contributes. McArdle emphasized two benefits of examining person log-likelihoods. First, he argued that examining person log-likelihood would help to clarify the interpretation of statistical indices, such as the likelihood-ratio-test. Second, he argued that examining person log-likelihoods would provide a means for examining the data for outliers or subgroups.

In the context of structural equation modeling, Reise and Widaman (1999) argued that PLL_i could quantify person fit, such that small negative values of PLL_i would indicate that a pattern of observed data was likely to occur given the model of interest, while large negative values would indicate that a pattern of observed data was not likely given the model of interest. Large negative values of PLL_i would therefore be evidence that an observation is an outlier given the model of interest. Because there were no means to compare PLL_i to a standardized distribution, Reise and Widaman (1999) used the formula in Equation 14, to calculate IND_CHI_i, an individual's contribution to overall model misfit

$$IND\_CHI_i = -2(PLL_{i_{sat}} - PLL_{i_{int}}) \qquad (15)$$

where $PLL_{i_{int}}$ is the person log-likelihood for the model of interest and $PLL_{i_{sat}}$ is the person log-likelihood for a saturated model. They argued that the IND_CHI_i would result in value that would represent an individual's contribution to the overall model fit chi-square. Thus, high positive values of IND_{CHI} would represent individuals who are making a large contribution to the overall model misfit. For

both real and simulated data, they found that the IND_CHI_i was normally distributed with a mean of zero and a standard deviation of 1.2 for real data and .2 for simulated data. Given the differences from the standard normal distribution, they recommended two strategies that researchers could use for evaluating IND_CHI_i. The first approach would be to conduct a simulation and judge any IND_CHI_i relative to the distribution from the simulation results. The second approach would be to select a criterion by which to select individuals for further examination, say the highest 1%. To evaluate the performance of IND_CHI_i, Reise and Widaman compared it to Z_l (Drasgow, Levine & Williams, 1985), a traditional person fit index from item response theory. They found that the two indices were only weakly correlated, but IND_CHI_i identified similar individuals as outliers as those identified by Z_l, suggesting that the IND_CHI_i could be an appropriate approach to evaluating person fit in structural equation models.

Coffman and Millsap (2006) also applied the work by Lange, Westlake, & Spence (1976) to decompose overall log-likelihood into a person-specific log-likelihood, which they describe as an "index of the covariance weighted distance between an individual's scores and the group averages (p 6)." Coffman and Millsap used the person log-likelihood to show that poor global model fit may hide good approximation to individual observations. They compared -2PLL_i to IND_CHI_i for evaluating individual model fit in LGCMs using simulated data from an unconditional linear LGCM. The results of the simulation study indicated that -2PLL_i and IND_CHI_i were only correlated .159 for the linear LGCM, suggesting that the indices provide different information regarding person fit. Examination of the -2PLL_i and IND_CHI_i showed that the largest -2PLL_i values indicated large deviations from a linear trajectory, while small -2PLL_i indicated trajectories that were more linear. Unlike the -2PLL_i, large IND_CHI_i did not indicate observations that deviated most from a linear trajectory and small IND_CHI_i did not identify linear trajectories. They also followed the recommendation of McArdle (1997) and examined the percent contribution to overall model misfit for individuals with high -2PLL_i. They found in their empirical example that even observations with the largest -2PLL_i did not contribute substantially to overall model misfit (.41%). This finding suggests that percent contribution to overall model misfit may not appear to be

very large for individuals with large -2PLL_i. Coffman and Millsap argued that identifying person fit can allow researchers to (1) identify individuals for whom the model fit well or poorly, (2) identify outliers, or (3) use person fit indices for model diagnostics, similar to regression diagnostics. They argued that these three purposes for evaluating model fit are important because they can help to refine models, eliminate data entry errors, and identify model misspecification.

A limitation to all of the proposed approaches to evaluating person level information in a LGCM is that very little work has been done to evaluate the effectiveness and limitations of these approaches. Sideris (2006) is one exception. Sideris compared several approaches to identifying aberrant observations in the context of SEM. In his research, Sideris extended the work done by Reise and Widaman (1999) to create an index of person fit that allowed for missing data. This index, called adj_ll_i, is created by calculating a person log-likelihood that accounts for the missing data pattern and weights the person log-likelihood by the number of measured variables for person i. The adj_ll_i, is calculated as follows

$$adj\_ll_i = -\frac{1}{2p_i}\left[p_i \ln(2\pi) + \ln\left|\hat{\Sigma}_{yyi}\right| + \left(y_i - \hat{\mu}_y\right)' \hat{\Sigma}_{yyi}^{-1}\left(y_i - \hat{\mu}_y\right)\right] \quad (16)$$

where p_i is the number of measured variables for person i and $\hat{\Sigma}_{yyi}$ is the covariance matrix for the individuals with the same missing patterns as person i. Sideris used a simulation study to compare this index of person fit to a model free residual analysis (Bollen, 1987), the IND_CHI_i (Reise & Widaman, 1999), and a mixture modeling approach (Yung, 1997). The adj_ll_i consistently outperformed IND_CHI_i at identifying aberrant observations. Its performance was similar to the residual analysis, but had the advantage of being able to handle missing data. The mixture modeling approach was found to be ineffective for identifying aberrant observations. These results suggest that it is possible that a PLL_i for a model of interest could be used for identifying aberrant observations and that it is not necessary to use the IND_CHI_i which includes the extra calculation of comparing the PLL_i from the model of interest to the PLL_i for a saturated model.

Although Sideris (2006) made some important contributions to our understanding of several approaches for evaluating person level information, his research also had several limitations. First, it did not apply these approaches in a LGCM context. It is likely that the results found by Sideris would be similar if his study were applied to LGCM data, but there could be different results using a LGCM which has more model constraints relative to the factor analysis model used in Sideris' research. Second, Sideris only examined the effectiveness of three approaches. It would be useful to examine some of the other proposed approaches to determine which approach is most effective. Finally, Sideris focused solely on identifying aberrant observations due to extreme observations and therefore did not consider other types of aberrance.

Limitations of Previous Work on Person Level Analysis for LGCMs

Many approaches have been proposed to examine person level information for LGCMs. However, there are several limitations to the work done thus far. First, there is very little information regarding how well the approaches perform in practice. A few studies provide simulations demonstrating the ability of an approach to identify aberrant observations due to extreme observations (e.g. Bollen & Arminger 1991; Sideris, 2006), but these studies have not examined all of the proposed approaches or all types of aberrance.

Second, there is almost no information on how these approaches perform relative to each other. Only Sideris (2006) compared several approaches to detecting aberrant observations. Currently, it is not clear which of the proposed approaches is better for identifying different types of aberrant observations. Researchers could try all of the proposed approaches when evaluating person level information, but it would be useful to have some information regarding how best to use these approaches.

A third limitation of the previous work is that many of the proposed approaches provide little information on how best to implement the approach. For example, one proposed approach is to examine the distribution of factor score estimates for extreme values, however little information is provided on

what criteria should be used to identify extreme values. This problem also appears when trying to quantify person level model fit with the person log-likelihood indices or RMSR_i. Similarly, when trying to use any of these indices to identify an appropriate functional form, one would need some criteria to choose one functional form over another. Some of the person level analysis approaches compare the person level information to a standard distribution (e.g. Bollen and Arminger's residual analysis). Other approaches compare person level information to a fixed cut-off criteria (e.g. top 1%) or recommend using bootstrapping to generate a distribution to use for comparison (Reise and Widaman, 1999). Any work comparing these approaches to assessing person level information should carefully consider the criteria for making decisions. These decision making criteria are essential in order to use person level information to identify different types of aberrant observations.

Current Research

The current research aims to address the limitations of previous work by evaluating the performance of several approaches for using person level information to identify different types of aberrant individuals when researchers use LGCMs. There are two primary goals of the current research. The first goal is to identify how well several person level analysis approaches perform at identifying aberrant observations. Thus, this research provides information on which methods to use to identify aberrant individuals, as well as providing information on how to use the method. Specifically, this research examines using cut-off criteria to identify aberrant individuals. The second goal was to determine which circumstances enhance or impair the performance of these approaches. As a result, this research provides information on when these methods can be used.

To narrow the focus of the current research, only a subset of the types of aberrant observations and the potential person fit approaches were examined. This research focused on extreme trajectory, extreme variability, and functional form aberrance. It also focused on three factor score based approaches and two person log-likelihood based approaches. The factor score approaches I examined were factor score estimates by themselves, factor score residual analysis, and RMSR_i. The person log-likelihood

based approaches I examined were -2PLL_i and IND_CHI_i. In addition to limiting the person fit approaches examined, the current research also focused on a simple LGCM, because it was not clear what type of aberrant observations could be identified using these approaches or what conditions would impact their performance in a simple application of LGCMs. Thus, the results of this research provides information on how these approaches perform for a minimally complex model with the understanding that the results may be overly optimistic for more complex LGCMs.

Given previous research and the goals of this study, there are several hypotheses I wanted to test with the current research. I first present some general hypotheses that I thought would hold for all types of aberrant observations then I will present my hypotheses for specific types of aberrance. My first general hypothesis was that the approaches would do better when observed data are closer to the underlying latent trajectory. Specifically, the data are closer to the underlying latent trajectory when there is low time-specific variability and therefore high communality between the observations and the latent factors. Factor score estimates would do better because the factor score estimates are closer to the true data generating latent trajectory parameters. This would also influence any examination of the factor score estimates or statistics summarizing the relationship between the factor score estimate predicted observations and the observed data (e.g. residual analysis or RMSR_i) by providing estimates that are less influenced by time-specific variability. This should improve the ability of the factor score approaches to identify aberrant observations due to extreme trajectories, extreme variability, or a different functional form. Approaches based on PLL_i should also do better with higher communality because it should result in PLL_i indices which reflect model misfit rather than high time-specific variability.

My second general hypothesis was that more observation times (t) would result in improved performance at identifying aberrant observations of any type. Greater observation times would improve performance because there is more information with which to detect aberrant observations of any type. In addition, I would expect any approaches using the regression method to obtain factor score estimates to show more improvement in performance than Bartlett's based approaches because the regression method

estimates will have less shrinkage with more time points, which should result in less biased factor score estimates.

Third, I hypothesized that in contrast to having more observation times, having a larger sample size would not improve detection of aberrant observations of any type. More observations (n) may make it harder to recognize aberrant observations. Pek and MacCallum (2011) discuss how the case influence is moderated by sample size such that an influential case in a small sample will tend to appear more influential than the same case in a larger sample. For this study, this could mean that measures of extreme trajectory or variability aberrance are more useful when sample size is not extremely large. The PLL_i based approaches may be more susceptible to this problem than the factor score approaches because these approaches make the comparison between observations and overall fit. With many observations, it may not be as clear who is contributing the most to misfit. For functional form aberrance, I hypothesized that a larger sample size would not help identify an appropriate functional form for a given person, because more observations do not provide any more information with which to identify an appropriate functional form for person i. This should be true for both the factor score based approaches and the person log-likelihood based approaches.

The prior hypotheses focus on conditions that would improve or worsen the identification of any type of aberrant observation with any person level analysis approach. There are also conditions that should influence specific person level analysis approaches such that some approaches should perform better than others. Next, I present my hypotheses regarding specific person level analysis approaches. First, I describe my hypotheses regarding the detection of extreme trajectory aberrance. Next, I present my hypotheses regarding the detection of extreme variability aberrance and then I present my hypotheses regarding functional form aberrance.

Hypotheses Regarding Detection of Extreme Trajectory Aberrance

I had several hypotheses that are specific to the detection of extreme trajectory aberrance. First, I hypothesized that examining factor score estimates by themselves would be useful for identifying aberrant observations if the individuals are aberrant due to extreme trajectories, because factor score estimates represent estimates of an individual's latent trajectory. Second, because the $RMSR_i$ and the factor score residual analysis provide a summary of how close the data are to the predicted observations and do not compare the data relative to the average trajectory, I hypothesized that these approaches would perform worse with aberrant observations due to extreme trajectories.

Regarding the person log-likelihood based approaches, I hypothesized that the -2PLL_i would do better than the IND_CHI_i at identifying aberrant observations due to extreme trajectories, based on the work done by Sideris (2006) and by Coffman and Millsap (2006). Sideris (2006) found that his adj_ll_i did better than the IND_CHI_i at identifying individuals with extreme observations in CFA models. The current research will not be examining missing data, so the adj_ll_i will be equivalent to the -2PLL_i divided by the number of observations. Although Coffman and Millsap (2006) did not examine rates of detection of aberrant observations, they did examine what observations were identified with -2PLL_i and IND_CHI_i. They found that the IND_CHI_i did not identify observations that deviated most from the model implied average trajectory, whereas the largest -2PLL_i did identify observations with large deviations from the model implied average trajectory. These results suggest that the IND_CHI_i may not be the best approach for detecting aberrant observations, but that the -2PLL_i should be able to identify observations that deviate from the model implied average trajectory.

I also hypothesized that the -2PLL_i would do better than the factor score approaches at detecting extreme trajectory aberrance based on the findings in Coffman and Millsap (2006). Specifically, they found when they fit a linear LGCM and examined the observations with the largest -2PLL_i that the two observations with the largest -2PLL_i showed nonlinear trajectories, and one of them appeared to have a

substantially higher intercept than the model implied average trajectory. These results suggest that the -2PLL_i may out-perform all of the factor score approaches because it can identify observations with any type of deviation from the model implied average trajectory. Therefore, because it captures any deviations from the model implied trajectory, using the -2PLL_i may be more efficient than examining several factor score estimates to assess different types of extreme trajectory aberrance.

Hypotheses Regarding Detection of Extreme Variability Aberrance

I had several hypotheses that were specific to the detection of extreme variability aberrance. First, I hypothesized that examining factor score estimates by themselves would not detect individuals who are aberrant because of high variability in their observations. The average trajectory for individuals with extreme variability aberrance would not necessarily be extreme in terms of intercept or slope, so factor score estimates of the latent intercept and slope for each individual would not provide information that would help detect extreme variance aberrance.

Second, because the RMSR$_i$ and factor score residual analysis provide a summary of how close the data are to the predictions, I hypothesized that these approaches would provide a good estimate of how much time-specific variance a given individual has relative to the average individual. This information would help to identify observations that are aberrant due to extreme variability.

Third, I hypothesized that the both the -2PLL_i and the IND_CHI_i would do worse than both the RMSR$_i$ and factor score residual analysis at identifying aberrant observations due to extreme variability. Both the -2PLL_i and the IND_CHI_i focus on deviations from the model implied average trajectory, which should not provide as much information about time-specific variability as examining deviations from the individual model implied trajectory. However, I would expect that these approaches would capture more information on time-specific variability than examining factor score estimates by themselves.

Hypotheses Regarding Functional Form Aberrance

Given that there are no previous simulation studies that examine the use of person level information for identifying functional form aberrance, I had fewer hypotheses for how effective the person level information would be at identifying individuals with functional form aberrance, in this case identifying an appropriate functional form for an individual's latent trajectory. First, I hypothesized that the larger the difference between the data generation models, the easier it should be to identify an appropriate functional form for each individual in a given sample. For example, when Coffman and Millsap examined person level fit for a linear and quadratic model, they found that person level fit could indicate good fit for a linear model, even though the data were generated to be from a quadratic model. They hypothesized that this was because even though the quadratic component was statistically significant, the size of the quadratic effect was small, and therefore a linear model could adequately approximate the quadratic effect. It would therefore be reasonable to hypothesize that the degree to which normal and aberrant population data generating models are similar will impact the ability of person level information to distinguish individuals from one model over the other. Specifically, the greater the difference between the two models, the easier it should be for person level information to identify which functional form is the best representation of a given individual's underlying trajectory.

Second, I hypothesized that examining factor score estimates themselves would not provide an effective way to select a model. If many factor score estimates are near zero that may suggest that a factor is not necessary. However, it may be difficult to assess how close to zero a factor score estimate should be to suggest that a model with that factor is not part of the data generating model for a given individual.

Lastly, I hypothesized that person level approaches that evaluate how close a model is to the observed data should do better at identifying an appropriate functional form than examining factor score estimates. These methods would include factor score residual analysis, RMSR_i and $-2PLL_i$. To identify an appropriate functional form, one could examine the difference in the factor score residuals or RMSR_i for the models of interest. If the change in the factor score residuals or RMSR_i is small, one could argue

that the more complex functional form does not provide a substantial improvement in the model implied prediction for person i. Similarly, because the -2PLL_i captures an individual's deviation from the overall model implied trajectory, small changes in the -2PLL_i suggest that the more complex overall functional form does not lead to substantially lower model misfit attributed to person i.

Summary of Current Research

Given these hypotheses, I used three simulation studies to assess the performance of the factor score estimates, factor score residuals, RMSR$_i$, -2PLL_i, and IND_CHI_i. The first simulation study evaluated the ability of these approaches to detect extreme trajectory, while the second study examined the detection of extreme variability aberrance, and the third study assessed the ability of these approaches to identify functional form aberrance. A summary of each study is provided next.

The first simulation study utilized two models for generating individuals with extreme trajectory aberrance, and assessed the impact that communality, number of observation times, and sample size had on the detection of aberrant individuals. Ideally, all of the approaches should perform equally well regardless of the number of aberrant observations. To assess if this is true, the first simulation generated data such that either 2% or 10% of the data were aberrant. Since the data were known to come from the normal or aberrant data generating models, the performance of these approaches was measured by assessing true positive aberrant observations, sensitivity, and specificity. True positives are observations that are generated and identified as aberrant. Sensitivity assesses how well a given approach is at identifying aberrant observations. It is calculated as

$$sensitivity = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives} \quad (17)$$

where the number of false negatives is the number of individuals whose data were generated to be aberrant, but the approach identified as non-aberrant.

In contrast, specificity assesses how well a given approach identifies non-aberrant observations. It is calculated as

$$specificity = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives} \quad (18)$$

where the number of true negatives is the number of individuals whose observations were generated and identified as non-aberrant. The number of false positives is the number of individuals whose data were generated to be non-aberrant, but the approach identified as aberrant. Together, the true positive rate, sensitivity, and specificity of each approach were used to compare and contrast the ability of the approaches to identify extreme trajectory or variability aberrance.

The second simulation study assessed the performance of person level analysis approaches at detecting extreme variability aberrance. It examined the impact of the same experimental conditions as the first simulation study, including: communality, number of observation times, sample size, and proportion of aberrant observations. As with the first simulation study the true positive rate, sensitivity, and specificity were examined to determine which person level analysis approach works best at identifying extreme variability aberrance.

The third simulation study examined identifying functional form aberrance using data with observations generated from either a linear or quadratic latent growth curve population model. Unlike the first two simulation studies which examined the proportion of aberrant observations, this study examined the proportion of data generated from a quadratic model. Ideally, the person level information should be able to identify the correct data generating model for a given individual, regardless of how many people come from a given model. To evaluate this, each sample had some proportion of observations generated from a quadratic LGCM ranging from 10% to 40%. In addition, the magnitude of the quadratic effect was also examined by generating data to have a quadratic effect with either a smaller or larger size relative to the linear effect ($1/20^{th}$ or $1/10^{th}$ the size of the linear effect). As with the other simulation studies, the

third simulation study assessed the impact that communality, number of observation times, and sample size had on identifying the correct functional form data generating model.

Since the data were known to come from the linear or quadratic data generating models, the performance of the approaches with clear cut-off criteria for model selection were measured by assessing true positive model selection, sensitivity, and specificity. For approaches where the criterion for model selection was less clear, a Receiver Operating Characteristic (ROC) curve were used to evaluate different threshold criteria for model selection. An ROC curve is a plot of the proportion of people who are correctly classified as aberrant (true positive fraction, TPF) as a function of the proportion of people who are incorrectly classified as aberrant (false positive fraction, FPF). This approach shows the trade off in changing the criteria for classification such that higher TPF rates will also result in higher FPF rates. If the data generating model membership (in this case, linear vs. quadratic) is not related to the person level information, then classification would be arbitrary and the ROC curve would be a diagonal line dividing the plot in half. On the other hand, if the data generating model membership is related to person level information, then the area under ROC curve will be more than half the plot. The greater the area under the ROC curve, the greater the association between membership and person level information. One way to quantify the accuracy of the person level information at classification is to calculate the area under the curve (AUC). AUC ranges from 0 to 1.0, with .5 indicating that classification is equivalent to random classification and 1.0 indicating perfect classification. AUC was calculated for all of the approaches, in order to compare and contrast the ability of these person level information approaches to identify functional form aberrance.

Together, these simulation studies provide information on whether person level information can be used to identify different types of aberrant observations. In addition, these studies evaluate if cut-off criteria can be used with these approaches, as well as some information about what criteria to use. Lastly, by examining several conditions that may enhance or impair these approaches, this research provides information on what conditions are necessary to successfully implement these approaches.

Chapter 2

STUDY 1: Evaluating Approaches to Detect Extreme Trajectory Aberrance in LGCMs

This chapter examined several approaches for analyzing LGCMs to determine which approach or approaches were the best at identifying extreme trajectory aberrance. To evaluate the person level analysis approaches, data sets were simulated such that the data were generated from either a normal or an aberrant population. Next, a model based on the normal population was fit to the data, and a variety of approaches to identify extreme trajectory aberrance were applied. Specifically, this chapter examined five person level analysis approaches to identify extreme trajectory aberrance: (1) factor score estimates by themselves, (2) factor score residual analysis, (3) RMSR_i, (4) $-2PLL_i$, and (5) IND_CHI_i. To quantify the performance of these approaches, this chapter assessed the frequency at which data generated from the aberrant population were detected as aberrant, as well as the sensitivity and specificity of each approach.

Another goal of this chapter was to determine which conditions hinder or help approaches to correctly identify extreme trajectory aberrance. In Chapter 1, I hypothesized that higher communality and more observation times would improve detection of aberrant observations. To assess these hypotheses, data were generated to have either low (.4) or high (.8) communality, and either 4 or 8 observation times. To evaluate the hypothesis that higher sample sizes would not improve detection of aberrant observations, data were generated to have either 200 or 1000 observations. Lastly, in Chapter 1, I stated that ideally the performance of any approach should not depend on the proportion of aberrant observations. To examine this, data sets were generated to have either 2% or 10% of the observations from the aberrant data population.

In addition to these hypotheses, Chapter 1 also contained several hypotheses regarding how well different approaches would perform at identifying extreme trajectory aberrance. To examine these hypotheses in Chapter 2, data w   berrance in one of two ways.

For the first form of aberrance, data were generated to be from a population with extreme intercepts relative to the normal population. For the aberrant population, the data were generated from a population model with the same correlation between the intercept and slope, and the same slope mean and variance as the normal population. To obtain observations with extreme intercepts, the intercept mean was generated to be either 3 intercept standard deviations higher or lower than the intercept mean for the normal population. To make sure that the data were generated to have extreme intercepts, the variance for the aberrant population's intercepts was smaller than the normal population.

For the second form of extreme trajectory aberrance, data were generated to be from a population with both extreme intercepts and slopes relative to those of the normal population. To generate this data the observations from the aberrant population were generated to have a mean intercept and a mean slope that were either 3 standard deviations higher or lower than the means from the normal population. As with the extreme intercepts only aberrance condition, the extreme intercept and slope condition had smaller variances for the intercept and slope, while preserving the correlation between the intercept and slope.

Given the conditions that were used to test the hypotheses of interest, the simulation study in this chapter has 32 condition combinations. The simulation includes two models for generating aberrant observations: extreme intercepts, and extreme intercepts and slopes. There are two proportions of aberrance (2% or 10%), two sample sizes (N=200, 1000), two number of observation times (4 or 8), and two levels of communality (.4 or .8). The subsequent sections of Chapter 2 describe the population models, data generation, fitted model, approaches for identifying aberrant observations, methods for evaluating the diagnostic procedures, and results of the study.

Population models

The population values for the covariance matrix and mean vector for the latent trajectory parameters for the normal population come from McArdle and Bell (2000). The parameter estimates were obtained using a subset of data from the National Longitudinal Study of Youth (NLSY; Baker, Keck,

Mott, & Quinlan, 1993; Chase-Lansdale, Mott, Brooks-Gunn, & Phillips, 1991). The data include 233 children who participated in the NLSY in 1986 when the children were ages 6 to 8. The data were selected because this subset of children had measures of their reading ability 4 times over 8 years. McArdle and Bell provide more information on how the reading scores were calculated. This study used the linear LGCM parameter estimates presented in Table 5.4 on page 83 of McArdle and Bell (2000):

$$\boldsymbol{\Phi} = \begin{bmatrix} 74.0 & \\ 1.8 & 2.5 \end{bmatrix} \quad \boldsymbol{\mu}_{\eta} = \begin{bmatrix} 32.2 \\ 6.5 \end{bmatrix}$$

These parameters were used to generate the latent trajectory parameters for each person in the normal population. To ensure that no aberrant observations were included in the normal population, only latent trajectory parameters within 2 standard deviations of the mean parameters were included in the normal population.

To generate latent trajectory parameters for the aberrant population with extreme intercepts, the following covariance matrix was used:

$$\boldsymbol{\Phi}_{ext\_int} = \begin{bmatrix} 9.0 & \\ .6 & 2.5 \end{bmatrix}$$

and half of the aberrant observations were generated using the high intercept mean vector:

$$\boldsymbol{\mu}_{\eta\_highint} = \begin{bmatrix} 58.0 \\ 6.5 \end{bmatrix}$$

while the other half of the aberrant observations were generated using the low intercept mean vector:

$$\boldsymbol{\mu}_{\eta\_lowint} = \begin{bmatrix} 6.4 \\ 6.5 \end{bmatrix}$$

To generate the aberrant trajectory parameters for the data generated to have an aberrant population with both extreme intercepts and slopes, the following covariance matrix was used:

$$\boldsymbol{\Phi}_{ext\_int} = \begin{bmatrix} 9.0 & \\ .2 & .5 \end{bmatrix}$$

with a fourth of the aberrant observations generated from each of these mean vectors:

$$\boldsymbol{\mu}_{\eta_{bothhigh}} = \begin{bmatrix} 58.0 \\ 11.2 \end{bmatrix} \qquad \boldsymbol{\mu}_{\eta_{lowhigh}} = \begin{bmatrix} 6.4 \\ 11.2 \end{bmatrix} \qquad \boldsymbol{\mu}_{\eta_{highlow}} = \begin{bmatrix} 58.0 \\ 1.8 \end{bmatrix} \qquad \boldsymbol{\mu}_{\eta_{lowlow}} = \begin{bmatrix} 6.4 \\ 1.8 \end{bmatrix}$$

These population parameters were used to generate the latent trajectory parameters for the normal and aberrant populations. Next, I will describe how the observed data were generated to match the experimental conditions.

Data Generation

For each of the experimental conditions, 500 samples were generated to have either 200 or 1000 observations. Data were generated to have either 2% or 10% of the data be from the aberrant population. After the individual latent trajectory parameters ($\boldsymbol{\eta}_i$) were generated using the appropriate population model, a linear LGCM was used to transform the individual latent trajectory parameters into the observed data. Equation 1 was used to transform the individual latent trajectory parameters into the observed data ($\boldsymbol{y}_i$). For the conditions with 4 time points, the following factor loading matrix was used:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}$$

and for the conditions with 8 time points, the following factor loading matrix was used:

$$\boldsymbol{\Lambda} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \\ 1 & 6 \\ 1 & 7 \end{bmatrix}$$

Both of these matrices specify that the data follow a linear model with time 1 specified as the intercept.

For each observation generated from the normal population or the aberrant populations with extreme intercepts or slopes or both, a time-specific residual (ε_{it}) was added at each time point. The size of the variance of the time-specific residual ($VAR(\varepsilon_{it})$) depended on the target communality for a given data generating condition. This is because communality is a measure of how much variance in the observed data are explained by the latent factors, therefore, the conditions where the data were specified to have a communality target of .8 had a smaller time-specific error variance than the conditions with communality target of .4.

Fitted Model

A linear LGCM, as described in Equation 1, was fit to every data set from every data generating condition. The linear LGCM was fit in SAS 9.2 using SAS PROC CALIS (SAS Institute, 2008). The LGCM was parameterized to have the same lambda matrix specified as the one used to generate the observed data.

Identifying Aberrant Observations

Five approaches to identify extreme trajectory aberrance were applied to the results from the fitted models. The five approaches were implemented using SAS PROC IML. Several criteria for aberrant observations were examined and then true positive rate, sensitivity, and specificity were calculated for each approach in each data generating condition. The criteria for each approach will be described next.

For the first approach, I examined factor score estimates using scores generated from the regression and Bartlett's methods. The factor score estimates were calculated using Equation 5 and 6. For both methods, I identified any observations as aberrant if they were extreme observations such that they were in the top or bottom 1% of the data, if the data were generated to have 2% aberrant observations. Similarly, if the data were generated to have 10% aberrant observations, I identified the top and bottom 5% of the observations as aberrant. For the conditions where the data were generated to have extreme

intercepts only, the latent intercept factor was examined. For the conditions where both the intercept and slope factors are generated to be aberrant, both the intercept and slope were examined for aberrant observations. One additional criterion that was examined for these conditions is whether requiring observations to have both an aberrant intercept and slope to be considered aberrant would improve the detection of aberrant observations or if having either an aberrant intercept or slope would provide better detection.

For the second approach, examining factor score residual analysis, I first calculated factor score estimates using the regression and Bartlett's methods and then used Equation 8 and Equation 13 to calculate the residuals and the residual test statistic. As previously stated, the test statistic for the residuals follows a chi-distribution with degrees of freedom equal to the number of observed variables, in this case, the number of time points. In the conditions with observations generated to have 2% aberrance, a cut-off criterion from the relevant chi-square distribution was obtained such that 2% of the observations were beyond the cut-off criterion. A similar procedure was used for the conditions where 10% of the observations were generated to be aberrant. In addition to these cut-off criteria, I examined a second cut-off criterion that accounts for the issue of repeated significance testing by using a Bonferroni correction to the significance level. This correction may be too conservative and select too few observations, but in contrast to the traditional criteria provides information on the degree to which any cut-off criterion may need to be corrected to make use of the residual test statistic.

The third approach, RMSR_i, was calculated using Equation 9. Small positive values of RMSR_i indicate that the predicted values of the model are close to the observed values and large positive values of RMSR_i indicate a poor prediction of the observed data. Therefore, aberrant observations were identified by selecting the observations with RMSR_i values that were the highest 2% or 10% of the RMSR_i values within a given data set.

The fourth approach, $-2PLL_i$, was calculated by taking negative two times Equation 14. Similar to the $RMSR_i$, small positive values of $-2PLL_i$ indicate that the model is a good representation of the data and large positive values indicate that the model is a poor representation of the data. Thus, individuals were identified as aberrant by selecting observations with the highest 2% or 10% of the $-2PLL_i$ values within a given data set.

Lastly, the fifth approach, IND_CHI_i, was calculated using Equation 15. Large positive values for IND_CHI_i indicate that the model of interest is a poor representation of the data relative to the saturated model, while small positive or negative IND_CHI_i indicate that the linear model is a good representation of the data. Therefore observations with the highest 2% or 10% of IND_CHI_i values within a given data set were identified as the aberrant observations.

Evaluation of Diagnostic Procedures

As previously described, the performance of these approaches for identifying aberrant observations was quantified using true positive rate, sensitivity, and specificity. These calculations were made for each data set in the study. Within a given combination of experimental conditions, descriptive statistics including mean, median, minimum, and maximum were used to compare the approaches. To test the impact of levels within a condition, a series of ANOVA models were fit using the experimental conditions as factors explaining the performance indicators. The first ANOVA model tested the main effects of experimental condition, while the subsequent ANOVA models evaluated interactions among the experimental conditions. To assess which experimental conditions have the largest effects I examined the semipartial η^2 effect size estimates. Together, the descriptive statistics and ANOVA models were able to identify which approaches were able to accurately detect aberrant observations, and under what conditions the accuracy was better or worse.

Results

Overall Performance

Table 1 contains the mean number of true positives within each combination of percent aberrant observations and sample size for the conditions where the data in the aberrant population had extreme intercepts only. The results in Table 1 indicate that as hypothesized the methods that examine the factor score estimates directly identified the most aberrant observations on average followed by the -2PLL_i approach. The approaches using regression factor score estimates did slightly better than those using Bartlett's factor score estimates. The IND_CHI_i, RMSR$_i$ and the methods examining factor score estimate residual statistics (FS Residuals and Bonf. FS Residuals) identified the fewest true aberrant observations. The Bonferroni corrected factor score estimate residual method (Bonf. FS Residuals) detected almost no aberrant observations and therefore will not be examined in any of the ANOVA analyses.

Table 2 contains the mean number of true positives for the conditions where the data in the aberrant population were generated to have both an extreme intercept and slope. Overall the results are very similar to those found in Table 1. The approach examining either factor score estimate does the best, followed by the -2PLL_i approach and then the approach examining both factor score estimates. Again the approaches examining factor score residuals do poorly, especially the Bonferroni corrected factor score estimate residual method which also detected almost no aberrant observations. Unlike the condition where the data were generated to only have extreme intercepts, the results in Table 2 suggest that the -2PLL_i approach does better when the data were generated to have both an extreme intercept and slope. In this case the -2PLL_i approach does slightly better that the approach examining the factor score estimates directly when both the intercept and slope factor score estimates are required to be identified as aberrant. When comparing the methods for estimating factor score estimates, the regression factor score method tended to do better when the approach required both factor score estimates to be identified as extreme, whereas Bartlett's tended to do better with the approach requiring either factor score estimate to be

extreme. This suggests that the regression method does a better job of producing factor score estimates that preserve the extremity of both the intercept and slope factors. Overall none of the methods are able to identify all of the aberrant observations due to extreme trajectories, but the approaches examining factor score estimates directly and the -2PLL_i approach worked the best of the approaches examined.

Because there were a different number of potential true positives depending on the sample size and percent aberrant observations, it is more useful to examine sensitivity which is the ratio of true positives to the overall number of aberrant observations. Table 3 and Table 4 contain descriptive statistics for sensitivity across conditions. They include the mean, standard deviation, median, minimum, and maximum sensitivity. Table 3 shows the results for when the aberrant data were generated to only have extreme intercepts, while Table 4 contains the sensitivity results for the conditions where the aberrant data were generated to have extreme intercepts and slopes. The results in these tables show that examining the factor score estimates directly tends to identify two-thirds to three-fourths of the aberrant observations when examining the intercept factor score estimate alone or either factor score estimate. On average the sensitivity of the -2PLL_i approach was .32 when the aberrant data were generated to only have an aberrant intercept and .52 when both the intercept and slope were generated to be extreme. For both types of extreme trajectory aberrance the factor score residual approaches and IND_CHI_i identified on average less than 20% of the aberrant observations. The standard deviations in Table 3 and Table 4 are smallest for the factor score estimate residual methods, RMSR_i and IND_CHI_i, indicating that these methods had more consistent sensitivity across the conditions. The median sensitivity was similar to the mean for all the approaches, with the exception of the approach examining intercept Bartlett's factor score estimate directly in the condition where only the intercepts were generated to be aberrant. In this case the median was .85, while the mean was .76, suggesting that more than half of the sensitivity results were greater than the mean sensitivity for this approach. All the approaches had a minimum sensitivity of 0, meaning that for at least one simulated data set, none of the aberrant observations were correctly identified as aberrant. Finally, Table 3 and Table 4 show that the only approaches to have a maximum sensitivity of 1, meaning

that all of the aberrant observations were identified, were the approaches examining intercept factor score estimates (when only the intercepts were generated to be aberrant) and examining either factor score estimate (when both trajectory parameters were generated to be extreme), as well as the -2PLL_i approach for both types of extreme trajectory aberrance.

In addition to examining sensitivity it is also useful to examine specificity which in this case measures how frequently the non-aberrant observations are identified as non-aberrant. Table 5 and Table 6 contain the mean, standard deviation, median, minimum, and maximum specificity across conditions. It is clear from Table 5 and 6 that on average all of the approaches have high specificity, typically 95% or more of the non-aberrant observations are identified as non-aberrant. After excluding the Bonferroni corrected factor score residual approaches (due to their poor sensitivity), the approaches with the highest specificity were the factor score approaches examining both factor score estimates (when both trajectory parameters were generated to be aberrant) and the intercept factor score estimate approach (when only the intercept was generated to be aberrant), followed by Bartlett's factor score estimate residual approach and the -2PLL_i approach. The standard deviations in Table 5 and Table 6 are small for all of the approaches indicating that specificity was consistent across the conditions. The median specificity was typically higher than the mean for all the approaches, typically around .97. Minimum specificity ranged from .83 for the approach examining the intercept regression factor score estimate to .99 for the Bonferroni corrected factor score residual analysis. These minimums suggest that the approaches have good specificity across conditions. Finally, Table 5 and 6 shows that most of the approaches have a maximum specificity of 1, meaning that all of the non-aberrant observations were identified as non-aberrant for at least one data set.

This high specificity is likely an artifact of the criteria used to identify observations as aberrant or non-aberrant. Given that the criterion for most of the approaches requires that 90% or 98% of the data in a given data set be identified as normal, we would expect sensitivity to be high. In the conditions where 2% of the observations were generated to be aberrant and the cut-off criterion requires 98% of the

observations to be identified as non-aberrant, by chance the specificity would be .96. Similarly for the conditions where 10% of the observations were generated to be aberrant and the cut-off criterion requires 90% of the observations to be identified as non-aberrant, by chance the specificity would be .81. Since this study has balanced conditions we would therefore expect the sensitivity to be .885 by chance alone. The sensitivity is higher than this on average, suggesting that the approaches are doing better than chance at identifying non-aberrant observations.

The results in Tables 1 to 6 suggest that overall the approaches were highly specific, but not highly sensitive. As hypothesized the factor score approaches that examine the calculated factor score estimates directly did a better job of identifying extreme trajectory aberrance than the residual based methods. Tables 1 to 6 also show that as hypothesized the -$2PLL_i$ outperformed the IND_CHI_i in both sensitivity and specificity. This suggests that the -$2PLL_i$ should be favored over the IND_CHI_i when trying to identify extreme trajectory aberrance. Compared to the factor score approaches, the -$2PLL_i$ did not perform best overall in terms of sensitivity, but it had higher specificity than the approach examining either factor score estimate, which was the most sensitive approach when the aberrant data were generated to have both extreme intercepts and slopes.

Impact of Experimental Conditions

To test the hypotheses regarding the impact of communality, number of observations, sample size, and proportion of aberrant observations several ANOVA models were examined. Because the number of potential true positives depends directly on the sample size and proportion of aberrant observations, it is more informative to examine sensitivity rather than number of true positives. Therefore, the ANOVA models were only used to determine the impact of the experimental conditions on sensitivity and specificity. The first ANOVA model examined the main effects of the experimental conditions, and subsequent models tested for interactions among the experimental conditions. The first interaction model examined all possible two-way, three-way and four-way interactions. Higher order interactions were then removed until the remaining interactions accounted for at least 1% of the total variance explained by the

model (η^2>=.01). The main effects and interaction models were applied to each method (excluding the Bonferroni corrected factor score estimate residual approach) for both sensitivity and specificity.

Tables 7 to 11 contain the final sensitivity ANOVA model results for each approach. Table 7 contains the results for the approaches examining the intercept factor score estimate directly and the factor score estimate residual based approaches for the sensitivity results from when the data were generated to have extreme intercepts only. Table 8 and Table 9 contain the results from when the aberrant data were generated to have both extreme intercepts and slopes. Table 10 contains the results for the log-likelihood based approaches for the conditions where the only the intercepts were generated to be aberrant and Table 11 contains the results for the log-likelihood based approaches for the conditions when both the intercepts and slopes were generated to be aberrant.

The results in Tables 7 to 11 show that sample size explained little (η^2=.01) or almost none (η^2<.01) of the variance in sensitivity for all the approaches. There were also no interactions with sample size found with effect sizes greater than .01. These results provide some evidence that the hypothesized null effect of sample size may hold when applying these approaches to the detection of extreme trajectory aberrance.

The size of the effect of the number of observation times varied from no effect (η^2<.01) to a moderate effect (η^2>=.27) depending on the approach and several interactions between number of observations and either communality or the percent of aberrant observations were found. Table 7 shows that for the approaches examining intercept factor score estimates, number of observation times interacted with communality for both the regression method (η^2=.12) and Bartlett's method (η^2>=.19). Figures 1 and 2 show the mean sensitivity of these approaches as a function of number of observation times and communality. These figures indicate that higher communality and number of observation times is associated with higher mean sensitivity, but that the effect of number of observation times is much larger when the communality is low and there is almost no difference when the communality is high. This suggests that as long as researchers have high communality the number of observation times is not as

important, but if a researcher has low communality, having more observation times can result in higher sensitivity. There was also a small interaction effect for the interaction between number of observation times and percent of aberrant observations (η^2=.01). Figure 3 contains the mean sensitivity by the number of observation times and percent of aberrant observations. It shows that a higher number of observation times and percent of aberrant observations are both associated with greater specificity, but that at lower observation times the effect of the percent of aberrant observations is slightly greater.

Table 8 shows that interactions with number of observation times were also found for the approaches examining factor score estimates directly for the conditions where the aberrant data were generated to have both extreme intercepts and slopes. A three-way interaction between number of observation times, communality, and percent of aberrant observations was found for the approach requiring both the regression factor score estimates to be aberrant (η^2=.02). Figure 4 shows mean sensitivity as a function of the conditions in this three-way interaction. It shows that for all combinations of number of observation times, communality, and percent of aberrant observations, the mean sensitivity of the approach examining regression factor score estimates is about .45, with the exceptions of the conditions where all three factors are low or all three factors are high. When all three factors are low the mean sensitivity is about .3 and when all three are high the mean sensitivity is about .7. These results suggest that only when these conditions all high or low do they influence the average sensitivity of this approach.

In addition to this interaction, Table 8 shows that for the approach examining both of the Bartlett's factor score estimates there was an interaction between number of observations and communality (η^2=.03). Figure 5 shows that this interaction is significant because both higher communality and higher number of observation times are associated with greater sensitivity, but that the effect of number of observation times is greater at higher communality. The results in Figure 5 show that the approach examining both Bartlett's factor score estimates is only effective at detecting extreme intercept and slope aberrance if a researcher has both high communality and high observation times. An interaction between these conditions was found for the approach examining either regression factor score

estimate (η^2=.06). Figure 6 shows that as previously found, higher communality and higher number of observation times are associated with greater sensitivity, but that there is almost no effect of number of observation times when communality is low. These results suggest that for the approach examining either regression factor score estimate having more observation times is more helpful when there is higher communality. Lastly, Table 8 indicates that there was a main effect of number of observation times on sensitivity for the approach examining either Bartlett's factor score estimate (η^2=.11). Examining the within condition means in Table 8 shows that more observation times were associated with greater sensitivity.

An examination of Table 7 and 9 shows that there was no effect of the number of observation times on sensitivity for most of the factor score residual based approaches. The only exceptions were some small interaction effects. When the aberrant data were generated to only have extreme intercepts the approach examining Bartlett's factor score residual analysis had a small interaction effect for the interaction between the number of observation times and the percent of aberrant observations(η^2=.01). Figure 7 shows the mean sensitivity for the Bartlett's factor score estimate residual approach is almost zero for all combinations of these conditions, but that having a higher percent of aberrant observations is associated with higher mean sensitivity and the effect is even greater at a higher number of observation times. Table 9 shows a similar interaction for Bartlett's factor score estimate residual approach in the conditions where the aberrant data were generated to have both extreme intercepts and slopes (η^2=.01) and Figure 8 contains a similar pattern of results to those found in Figure7.

Table 9 also shows small interactions between number of observation times and communality for both the regression factor score estimate residual approach (η^2=.03) and regression RMSR_i (η^2=.01). Figures 9 and 10 shows that for both of these approaches there is almost no effect of observation times at low communality and a negative effect of observation times at high communality. These results suggest that for the approaches using regression factor score residual analysis observations that are aberrant due to extreme trajectories are even less likely to be identified when the data are both closer to the underlying model (higher communality) and there is more information about a given person (more observation

times). This is counter to my hypotheses for these conditions, but makes sense in this case because the factor score residual analysis focus on how far a person's observations are from his or her model implied trajectory and not the overall trajectory. With greater communality and more observation times, these approaches should be identifying more people with greater variability around their observed trajectory and not greater distance from the model implied average trajectory.

For the person log-likelihood based approaches, Table 10 and Table 11 show small or no effect of the number of observation times. For the -2PLL_i approach, the number of observation times had a small interaction with communality (η^2=.04) when the aberrant data were generated to have extreme intercepts only and had no effect when the aberrant data were generated to have both extreme intercepts and slopes. Figure 11 shows that the interaction between number of observation times and communality for the -2PLL_i approach which is similar to the interaction effects shown in Figure 9 and 10. Figure 11 shows that the on average the sensitivity for the -2PLL_i approach is around .3, but at high communality and low observation times the sensitivity is greater, around .4. This seems to suggest that when the aberrant data are generated to only have extreme intercepts the -2PLL_i approach does best when the data are closer to the underlying trajectory and there are less observation times for a given person. This may occur because when there are more time points, the -2PLL_i approach may identify more individuals as aberrant because of other unusual data patterns. For the IND_CHI_i approach, Table 10 and 11 indicates a small effect of number of observation times such that more observation times were associated with lower sensitivity.

I had hypothesized that more observation times would be associated with greater sensitivity. The results for number of observation times indicate that for the approaches examining the factor score estimates directly, having more observations resulted in higher sensitivity providing some support for my hypothesis. However, for the other approaches the number of observation times either had no effect or more observations resulted in slightly lower specificity. Together these results suggest that number of observation times is related to the performance of the approaches examining factor score estimates directly, but does not seem to influence the performance of the other approaches.

Similar to the number of observation times, the effect of communality on sensitivity varied by the approach and the type of extreme trajectory aberrance. For the approaches that examine factor score estimates directly greater communality was found to have a meaningful main effect on sensitivity (η^2ranging from .14 to .44), such that higher communality was associated with greater sensitivity. In addition to these many effects there were also several interactions with number of observations found (see Figures 1, 2, 4 and 5). In addition to these interactions there was also a small interaction with percent of aberrant observations found when examining either Bartlett's factor score estimate when the aberrant data were generated to have both extreme intercepts and slopes. Figure 12 shows that that higher levels of percent of aberrant observations and communality were associated with higher sensitivity, but the effect of aberrant observations was smaller at the higher communality level. This seems to suggest that with greater communality the sensitivity of the approach examining either Bartlett's factor score estimates does not depend as much on how many aberrant observations are in the data for aberrant observations due to extreme intercepts and slopes.

For the factor score residual approaches, Table 7 and 9 indicate no or little effect of communality on the sensitivity of these approaches. Small interaction effects between communality and number of observation times were found for the regression factor score estimate residual approach (η^2=.03) and regression RMSR_i (η^2=.01). As previously described Figures 9 and 10 show these interactions. Overall, these results suggest that communality has little impact on the sensitivity of the factor score residual based approaches.

For the person log-likelihood based approaches, Table 10 and 11 show there was little or no effect of communality on the sensitivity of the IND_CHI_i approach. For the -$2PLL_i$ approach, there was a small interaction between communality and number of observation times (see Figure 11) when the aberrant data were generated to have extreme intercepts only. Table 11 shows that communality had a large main effect on the sensitivity of the -$2PLL_i$ approach (η^2=.46) when the aberrant data were generated to have both extreme intercepts and slopes. The within-condition means indicate an average

sensitivity of .42 when the communality was .4 and an average sensitivity of .63 when the communality was .8. These results suggest that communality had a large impact on the ability of the -2PLL_i approach to detect aberrance due to extreme trajectories when both trajectory parameters were extreme.

The communality results for the approaches that had the best average sensitivity tended to support the hypothesis that greater communality would be associated with better detection of aberrant observations. The approaches with the worse overall sensitivity typically showed no effect of communality on sensitivity. Therefore, for the approaches best suited to detect extreme trajectory aberrance, namely the approaches examining factor score estimates directly and the -2PLL_i approach, higher communality is associated with better detection of aberrant observations.

Lastly, Tables 7 to 11 show that the percent of aberrant observations had either main effect or interaction effects for all the approaches examined. Examining the within-condition means and previously described interaction effects revealed that a higher percent of aberrant observations was associated with a higher sensitivity. These effects are probably an artifact of the criteria used to select aberrant observations, because more observations would necessarily be selected as aberrant using the criteria I specified when more aberrant observations were in the data.

Together these results provide some support for the hypotheses that higher sample sizes would not lead to greater sensitivity, and that a higher number of time points and higher communality would be associated with greater sensitivity. More specifically, there is evidence for these relationships for the approaches with the greatest average sensitivity for detecting extreme trajectories, the approaches examining factor score estimates directly and the -2PLL_i approach. Although it was hypothesized that observation times would have a greater impact on sensitivity for regression factor score estimates, the results suggest that there was no difference or for the approaches examining factor score estimates directly there was evidence of the reverse, a greater effect size for observation times for the approaches using Bartlett's factor score estimates. One unsettling aspect of these results is that there was an effect of the percentage of outliers, such that having more outliers in a data set was associated with greater

sensitivity. This is troubling because one would want these approaches to work no matter how many aberrant observations are present in the data. It is likely that this finding is an artifact of the criteria selected for identifying aberrant observation, but further research on criteria for select aberrant observations would be needed to determine that this effect is only an artifact.

Tables 12 to 16 contain the final specificity ANOVA model results for each approach. Table 12 contains the results for the methods examining factor score estimates directly and factor score residual approaches when the aberrant data were generate to have extreme intercepts only. Table 13 and Table 14 contain similar results for when the aberrant data were generated to have both extreme intercepts and slope. Table 15 and 16 contains the results for the log-likelihood approaches.

The results in Table 12 to 16 provide further support for the hypothesis that sample size would be unrelated to the performance of these approaches. For all of the approaches examined there was no effect of sample size on specificity. For the approaches examining the factor score estimates directly there was typically a three-way interaction among the other factors: number of observation times, communality, and percent of aberrant observations. Figures 13, 14, 15 and 16 show these three-way interactions. These figures show a slight negative effect of the percent of aberrant observations on specificity, that is the greatest when there are fewer observation times and lower communality. More importantly all of the plots suggest that there is high specificity regardless of the levels of these conditions.

In the conditions where the aberrant data were generated to both have extreme intercepts and slopes, the approach examining both regression factor score estimates and the approach examining either Bartlett's scores did not have this three-way interaction. Instead Table 13 shows that the approach examining both regression factor score estimates had a meaningful interaction between the number of observation times and the communality (η^2=.03) and an interaction between communality and the percent of aberrant observations (η^2=.06). Figure 17 shows the interaction between number of observation times and communality is such that at low communality and a high number of observation times specificity is the lowest. However, this interaction effect is very small, such that the specificity is essentially the same for all combinations of these conditions. Figure 18 shows that the interaction between communality and

percent of aberrant observations and similar to Figure 17, the specificity is essentially equal for all combinations of these conditions, but is its lowest with low communality and a high percent of aberrant observations.

Table 13 also shows the approach examining either Bartlett's factor score estimates had a meaningful interaction between the number of observation times and the percent of aberrant observations (η^2=.03) and an interaction between communality and the percent of aberrant observations (η^2=.06). Figure 19 shows the interaction between number of observation times and the percent of aberrant observations is such that a greater percent of aberrant observations was associated with lower specificity and this effect was larger with a lower number of observation times. Figure 20 shows that the interaction between communality and percent of aberrant observations which has a similar pattern to the results in Figure 19, such that the negative effect of the percent of aberrant observations on specificity was greater at lower communality. Together the results for the approaches examining the factor score estimates directly suggest that the effects of these conditions on specificity are small and tend to occur at low levels of communality and number of observation times.

For both types of extreme trajectory aberrance, the specificity of the approach examining Bartlett's factor score residual analysis was found to be unrelated to sample size or communality, but a small interaction effect was found between number of observation times and percent of aberrant observations. Figure 21 and 22 show these interaction effects. As with the previous interactions, the differences in specificity tend to be small, with higher percent of aberrant observations being associated with lower specificity. Unlike the previous interactions, this effect tends to be slightly greater with more observation times for these residual based approaches.

For the regression factor score estimate residual approach, the RMSR_i and IND_CHI_i approaches there were no effects of any of the experimental conditions except the percent of aberrant observations. This suggests that any variability in specificity could be explained by the percentage of aberrant observations, and more likely that any variability in specificity was due to the criteria used to select aberrant observations.

For the -2PLL$_i$ approach when the aberrant data were generated to have extreme intercepts only, the percent of aberrant observations explained nearly all of the variability in specificity (η^2=.94), such that a higher percent of aberrant observations was associated with a lower specificity. There was also a small effect of number of observation times (η^2=.01), although the within condition means show no difference. When the aberrant data were generated to have both extreme intercepts and slopes, a similar effect was found for the percent of aberrant observation (η^2=.83), but there was no effect for number observation times, instead there was a small interaction between communality and percent of aberrant observations (η^2=.03). Figure 23 shows that the interaction between communality and percent of aberrant observations. Similar to the interactions found for the approaches examining factor score estimates directly, there was a negative effect of aberrant observations that was slightly greater in the low communality conditions. These results suggest that the primary factor explaining any variability in specificity for the -2PLL$_i$ approach was the percent of aberrant observation, which is likely an artifact from the criteria used to select aberrant observations.

Overall the specificity was high for all of the approaches. The ANOVA results provide support for the hypotheses regarding sample size, but almost no support for the hypothesis regarding number of observation times and communality. For all of the approaches examined the percent of aberrant observations explained the majority of the variability in specificity suggesting that most of the variability in specificity was an artifact of the criteria used to select non-aberrant observations.

Together with the results for sensitivity there is evidence that the performance of these approaches for detecting extreme trajectory aberrance is dependent on the number of aberrant observations such that more observations are identified as aberrant (regardless of whether they are truly aberrant) when there are more aberrant observations in the data. However, it is likely that this result is at least partially an artifact of the criteria used to select the aberrant observations. Evidence was found for the hypothesis that sample size is unrelated to the detection of extreme trajectory aberrance with these approaches. Similarly there was evidence to support the hypothesis that greater communality would be

associated with better detection of extreme trajectory aberrance with the largest impact found for the sensitivity of the approaches that examine factor score estimates directly. Unlike the hypothesized effect for number of observation times, there was only evidence that having more observation times has a small effect on the sensitivity for the approaches examining the factor score estimates directly. For the other approaches, observation times seem to be unrelated to identifying extreme trajectory aberrance.

Based on the average sensitivity and specificity, the best approaches for identifying extreme trajectory aberrance were the approaches examining the regression factor score estimates directly, followed by the $-2PLL_i$ approach. As one might expect, the conditions where both the intercept and slope where generated to be aberrant examining either factor score estimate was more sensitive, but requiring both factor score estimates to be extreme was the most specific approach. The $-2PLL_i$ approach was less sensitive than the approach examining both regression factor score estimates and more specific than the examining either regression factor score estimates (the most sensitive approach). These results suggest that if researchers are more concerned with identifying aberrant observations due to extreme trajectories correctly then they should examine either of the regression factor score estimates. If researchers want to balance identifying aberrant observations with not falsely identifying an individual as aberrant, they may prefer the $-2PLL_i$ approach or examining both regression factor score estimates. If researchers choose to use the regression factor score estimates, they should be aware that the sensitivity will depend on the number of observation times, communality and percentage of aberrant observations, such that higher levels of these variables will be associated with greater sensitivity. Researchers should also be aware that the specificity of the regression factor score estimates will depend on communality and percentage of aberrant observations, such that higher levels of communality are associated with higher specificity and a higher percentage of aberrant observations is associated with lower specificity. If researchers decide to use $-2PLL_i$, then they should be aware that higher communality is associated with higher sensitivity and specificity, while a higher percentage of outliers is associated with higher sensitivity and lower specificity.

Chapter 3

STUDY 2: Evaluating Approaches to Detect Extreme Variance Aberrance in LGCMs

This chapter examines several approaches for identifying extreme variability aberrance in LGCMs in order to determine which approach or approaches are the best at identifying observations that are aberrant due to high time-specific variability relative to the average time-specific variability. As with the previous study, I simulated data sets where the data are generated from either a normal or an aberrant population. Next, a model based on the normal population was fit to the data, and a variety of approaches for identifying extreme variability aberrance were applied. This chapter examines five approaches for identifying extreme variability aberrance, (1) factor score estimates by themselves, (2) factor score residual analysis, (3) RMSR_i, (4) $-2PLL_i$, and (5) IND_CHI_i. As with the previous study, the performance of these approaches for identifying extreme variability aberrance were quantified by assessing the frequency at which data generated from the extreme variability population were detected as aberrant, as well as the sensitivity and specificity of each approach.

To better understand what conditions hinder or help approaches to correctly identifying extreme variability aberrance, this chapter assesses several hypotheses presented in Chapter 1. First, I hypothesized that higher communality and more observation times would improve detection of aberrant observations. To assess these hypotheses, data were generated to have either low (.4) or high (.8) communality, and either 4 or 8 observation times. Second, I hypothesized that higher sample sizes would not improve detection of aberrant observations. To test this hypothesis data were generated to have either 200 or 1000 observations. Third, I stated that ideally the performance of any approach should not depend on the proportion of aberrant observations. To examine this, data sets were generated to have either 2% or 10% of the observations from the aberrant data population.

In addition to these hypotheses, Chapter 1 also contained several hypotheses regarding how well different approaches would perform with extreme variability aberrance. To assess the performance of several person level analysis approaches at detecting extreme variability aberrance, data were generated to be aberrant due to large time-specific errors either added or subtracted from each observation. Specifically, half of the observations for a given individual had 3 standard deviations of the time-specific error from the normal population added to their time-specific observation and half had 3 standard deviations subtracted from their time-specific observation. This resulted in higher time-specific variability for the observations for the individuals in the aberrant population.

Given the conditions that were used to test the hypotheses of interest, the simulation study in this chapter has 16 condition combinations. The simulation included two proportions of aberrance (2% or 10%), two sample sizes (*N*=200, 1000), two number of observation times (4 or 8), and two levels of communality (.4 or .8). The subsequent sections of Chapter 3 will describe the population models, data generation, fitted model, approaches for identifying extreme variability aberrance, methods for evaluating the approaches, and results of the study.

Population models

As with the previous study described in Chapter 2, the population values for the covariance matrix and mean vector for the latent trajectory parameters for the normal population, come from McArdle and Bell (2000). Details regarding the population parameters are described in Chapter 2. For this study, the normal population parameter estimates were used to generate the latent intercept and slope parameters for both the normal and aberrant population.

To generate data from the aberrant population due to extreme variability aberrance the observations from the aberrant population had larger time-specific error added to their observations at each time point. For the half of the aberrant population's observations, 3 time-specific error standard deviations were added to all of the odd time-points, and 3 error standard deviations were subtracted from

the even time-points. For the other half of the aberrant population 3 time-specific error standard deviations were subtracted from all of the odd time-points, and 3 error standard deviations were added to the even time-points. More details on this size of this standard deviation and details on how data were generated to match the experimental conditions of interest are described next.

Data Generation

Similar to the previous study, for each combination of the experimental conditions 500 samples were generated to have either 200 or 1000 observations. Data were generated to have either 2% or 10% of the data be from the aberrant population. After the individual latent trajectory parameters ($\boldsymbol{\eta}_i$) were generated, a linear LGCM was used to transform the individual latent trajectory parameters into the observed data. For each observation generated from the normal population a time-specific residual (ε_{it}) was added at each time point. The size of the variance of the time-specific residual ($VAR(\varepsilon_{it})$) depended on the target communality for a given data generating condition.

The data generated to be from the extreme variability aberrant population followed a slightly different data generation procedure. For any observations generated to be from the extreme variability aberrant population the mean and covariance matrix for the normal population were used to generate the latent trajectory parameters and Equation 1 was used to generate the observed data. The difference in the data generation was that in order to generate aberrant data with extreme variability each time-specific error in the aberrant population had three times the standard deviation from the normal population either added or subtracted from each time specific observation.

Fitted Model

As with the previous study, a linear LGCM, as described in Equation 1 was fit to every data set from every data generating condition. The linear LGCM were fit in SAS 9.2 using SAS PROC CALIS (SAS Institute, 2008). The LGCM was parameterized to have the same lambda matrix specified as the one used to generate the observed data.

Identifying Aberrant Observations

In an approach similar to that used in Chapter 2, five approaches for identifying aberrant observations were applied to the results from the fitted models. The five approaches will be implemented using SAS PROC IML. Several criteria for aberrant observations were examined for each approach and then the number of true positives, sensitivity, and specificity were calculated for each approach in each data generating condition. The criteria for the factor score residual analysis, RMSR_i, $-2PLL_i$, and IND_CHI_i, were the same as the criteria described in Chapter 2. For the approach using factor score estimates, I examined factor score estimates from both the regression and Bartlett's method. Unlike Chapter 2, for the approaches examining factor score estimates directly instead of just checking the intercept for aberrant observations, both the intercept and slope were examined. The same cut-off criteria described in Chapter 2 were used and I examined whether requiring observations to have both an aberrant intercept and slope or either aberrant would improve the detection of aberrant observations.

Evaluation of Diagnostic Procedures

As previously described, the performance of these approaches for identifying aberrant observations were quantified using true positive rate, sensitivity, and specificity. I followed the same procedure described in Chapter 2 such that these calculations will be made for each data set in the study and descriptive statistics, like mean, median, minimum, and maximum, were used to compare the approaches. The impact of levels within a condition were tested using a series of ANOVA models with the experimental conditions as factors explaining the performance indicators. The first ANOVA model tested the main effects of experimental condition, while the subsequent ANOVA models tested for interactions among the experimental conditions. The experimental conditions with the largest effects were determined using semipartial η^2 effect size estimates. Together, the descriptive statistics and ANOVA models, helped to identify which approaches were able to accurately identify extreme variability aberrance, and under what conditions the accuracy is better or worse.

Results

Overall Performance

Table 17 contains the mean number of true positives within each combination of percent aberrant observations and sample size. Table 17 shows that as hypothesized the factor score estimate residual methods detected more aberrant observations with extreme variability aberrance than examining factor score estimates directly. Table 17 also indicates that the -2PLL_i approach identified the most aberrant observations, which did not support the hypothesis that -2PLL_i would perform worse than the factor score estimate residual based methods at identifying extreme variability aberrance. Unlike the results from the previous study the approaches using regression factor score estimates tended to do better than the approaches using Bartlett's scores for the residual based approaches, but Bartlett's factor score estimates did better when examining the factor score estimates directly. Unlike extreme trajectory aberrance, the mean number of true positives does not appear to depend as much on the percent of aberrant observations.

To put the number of true positives in the context of how many aberrant observations were in a data set, Table 18 contains descriptive statistics for sensitivity across conditions. Table 18 includes the mean, standard deviation, median, minimum, and maximum sensitivity for each approach. According to Table 18 the -2PLL_i approach had the highest average sensitivity, followed by regression factor score residual analysis and regression RMSR_i. The least sensitive approaches were IND_CHI_i and the approaches examining both factor score estimates. The standard deviations indicate that the most sensitive approaches were also the most consistent; meaning that across the data sets the sensitivity was consistently high for these approaches. The medians show that in at least half of the data sets the most sensitive approaches identified nearly all or all of the aberrant observations as aberrant. The minimums indicate that while most of approaches had at least one data set where none of the aberrant observations were correctly identified, the most sensitive approaches were able to identify 25% of the aberrant

observations. The maximums indicate that most of the approaches had at least one data set where all of the aberrant observations were correctly identified as aberrant.

Table 19 contains the results for specificity which in this case measures how frequently the non-aberrant observations are identified as non-aberrant. Table 19 contains the means, standard deviation, median, minimum, and maximum specificity across conditions. There was little or no difference in the performance the approaches when comparing regression and Bartlett's factor score estimates. The mean specificity ranged from .92 to 1 and all of the standard deviations were all less than or equal to .05 indicating that there was little variability in specificity across the conditions. This lack of variability in specificity is further shown in the high median, minimum and maximum specificity. Given the consistently high specificity of all of the approaches, it does not make sense to examine what factors explain the little variability in specificity and therefore specificity will not be examined with the ANOVA analyses. As with the previous study, it is likely that this high specificity is at least partially due to the criteria selected to identify aberrant and non-aberrant observations.

The results in Tables 17 to 19 suggest that all the methods were highly specific, but only the -2PLL_i, regression factor score residual analysis, and regression RMSR_i were highly sensitive. These results did not support the hypothesis that the factor score residual based methods would outperform the-2PLL_i, but did support the hypotheses that the -2PLL_i and factor score residual based approaches would do better than examining the factor score estimates directly at identifying extreme variability aberrance. As would be expected, the Bonferroni corrected factor score residual analysis had lower sensitivity, but they didn't show any difference in specificity, suggesting that the correction would not reduce the number of non-aberrant observations identified as aberrant. Given these results it does not seem necessary to use the correction and it will not be considered in the ANOVA analyses.

Impact of Experimental Conditions

To test the hypotheses regarding the impact of communality, number of observations, sample size, and proportion of aberrant observations several ANOVA models were examined. For this study the

ANOVA models were only used to determine the impact of the experimental conditions on sensitivity. As previously stated there was very little variability in specificity and therefore none of the ANOVA models were fit to specificity. The same analytic strategy from Chapter 2 was used to assess the impact of the experimental conditions on the sensitivity in this chapter.

Tables 20 and 21 contain the final sensitivity ANOVA model results for each approach. Table 20 contains the results for the methods examining factor score estimates directly, and for the factor score residual based approaches and Table 21 contains the results for the log-likelihood approaches. For the approaches examining both factor score estimates, Table 20 indicates that there was a small or no effect sample size (η^2<=.01), providing some evidence for the hypothesis that sample size would be unrelated to the performance of these approaches. There were main effects of number of observation times (η^2=.07-.09), communality (η^2=.14-.39), and percent of aberrant observations (η^2=.10-.17) such that more observation times, higher communality and lower percent of aberrant observations were associated with lower sensitivity. In addition to these main effects several small interactions were found for the regression factor score estimates. For the approach examining both regression factor score estimates there was an interaction between number of observation times and communality (η^2=.01), as well as an interaction between communality and percent of aberrant observations (η^2=.01). Figure 24 and Figure 25 show these interactions. Figure 24 shows that both higher communality and more observation times are associated with lower sensitivity and that the effect of number of observation times is greater at lower communality. Figure 25 shows the same negative effect of communality, with a positive effect of percent of aberrant observations, which is greater at lower communality. Together these figures provide evidence that with greater communality and observation times, this approach is identifying observations as aberrant due to sources other than extreme variability. Figure 25 also suggests that there is likely a small effect of the criteria for selecting aberrant observations. For the approaches examining either regression factor score estimates there was an interaction between number of observation times and communality (η^2=.01). Figure 26 shows this interaction is similar to the one shown in Figure 24, suggesting that with greater

communality and more observation times the approach examining either regression factor score estimate is probably identifying observations as aberrant due to factors other than extreme variability.

These results do not fit with the hypotheses that more observation times and higher communality would be associated with higher sensitivity. Instead higher levels of these factors are associated with lower sensitivity. These results probably occurred because higher levels of these factors are associated with better model estimates which will result in factor score estimates that better capture the underlying trajectory, but more accurately estimating the underlying trajectory will not help detect which observations have extreme variability aberrance. Similar to the previous study the results for the percent of aberrant observations are potentially the result of the criteria chosen to select aberrant observations

Table 20 also contains the sensitivity ANOVA results for the factor score residual based approaches. The results indicate no effect of sample size on sensitivity for any of the approaches, which fits with the hypothesis that higher sample sizes would not be associated with better performance. The condition that explained the most variance in sensitivity for the residual approaches was the number of observation times. Examining the means for number of observation times shows that as hypothesized higher observation times were associated with higher sensitivity. Similar to sample size there was no effect of communality for any of the factor score residual approaches, which did not support the hypothesis that higher communality would be associated with higher sensitivity. Lastly, there was no effect of the percent of aberrant observations on sensitivity for any of the approaches except the Bartlett's $RMSR_i$ approach, where a small interaction effect with number of observation times was found (η^2=.02). Figure 27 contains a plot of this interaction which indicates that there is almost no difference in sensitivity for the two levels of percent of aberrant observations when there are more observation times, but with lower number of observation times a greater percent of aberrant observations is associated with slightly greater sensitivity.

Table 21 contains the results for the sensitivity ANOVA models fit to the log-likelihood based approaches. For $-2PLL_i$ there were no meaningful effect sizes for sample size or percent of aberrant observations, but main effects size were found for number of observation times (η^2=.28), communality

(η^2=.04) and an interaction between these two factors (η^2=.03). Figure 28 shows overall this approach has high sensitivity across the levels of these conditions. It also shows that greater observation times are associated with greater sensitivity and higher communality is associated with slightly lower sensitivity, except when there are more observation times. This suggests that having more observation times has a greater effect on the sensitivity of this approach, such that more observation times results in higher sensitivity regardless of the level of communality. Overall the means in Table 21 and the plot in Figure 28 show that regardless of the levels of these factors the -$2PLL_i$ has high sensitivity.

The results for IND_CHI_i in Table 21 indicate that there was no effect for communality and two interactions among the other experimental factors. Figure 29 contains plots for both the interaction between observation times and sample size and Figure 30 contains the interaction between observation times and percentage of aberrant observations. Both plots show that a greater number of observation times were associated with lower sensitivity, and that higher levels of sample size and aberrant observations were associated with higher sensitivity. The plots show that the effect of sample size is larger at a higher number of observation times and higher percentage of outliers. These results do not support the hypotheses regarding sample size, observation times or communality, and they suggest that the performance of the IND_CHI_i approach depends on the percentage of aberrant observations. The overall level of sensitivity in the means in Table 21 and shown if Figure 29 and 30 show that the IND_CHI_i has very low sensitivity relative to the -$2PLL_i$, suggesting that the -$2PLL_i$ is a much better approach for assessing extreme variability aberrance.

The sensitivity ANOVA results for using these approaches to detect extreme variability aberrance were not consistent across approaches. Some support was found for the hypothesis that sample size would be unrelated to the detection of aberrant observations, but there were small effect sizes for the approaches examining the regression factor score estimates directly and for an interaction between sample size and observation times for the IND_CHI_i approach. Similarly some support was found for the hypothesis that higher observation times would be associated with higher sensitivity, although again this wasn't the case

for the approaches examining the factor score estimates directly or for the IND_CHI_i approach. Higher communality was typically associated with lower sensitivity, which did not support my hypothesis regarding communality. Lastly, the results for the percent of aberrant observations typically indicated that more aberrant observations were associated with higher sensitivity.

Together these results suggest that these factors do not operate the same way when detecting extreme trajectory and extreme variability aberrance. However, more importantly all of these approaches had higher sensitivity when detecting extreme variability aberrance, although different approaches did better at detecting extreme variability versus extreme trajectory aberrance. The only approach that seemed to do relatively well at both types of aberrance was the -2PLL_i.

Chapter 4

STUDY 3: Evaluating Approaches for Detecting Functional Form Aberrance in LGCMs

This chapter examines the use of person level information for the detection of functional form aberrance. The design of this study essentially assesses two potential uses for person level information. First, it assesses the ability of person level information to be used for identifying model misspecification, in this case the misspecification of the functional form of the latent trajectory. Second, it examines the ability of person level information to be used for model comparison, which may result in new approaches for model selection. The main goal of this chapter was to learn which approach or approaches are the best at identifying observations from different LGCMs. To achieve this goal, data sets were simulated where the data in a given data set were generated from a population model with either a linear or quadratic latent growth curve. Next, a linear and a quadratic LGCM were fit to each data set, and a variety of approaches for identifying the data generating model using person level information were applied. Specifically, this chapter examines four approaches for identifying observations with aberrant functional forms, (1) factor score estimates by themselves, (2) difference in factor score residuals, (3) difference in RMSR_i, and (4) difference in $-2PLL_i$. Several indicators were used to quantify the performance of these approaches to identifying the data generating model. For the approaches with clear cut-off criteria for model selection, performance was measured by assessing true positive model selection, sensitivity, and specificity. For the approaches where the criteria for model selection was less clear, a Receiver Operating Characteristic (ROC) curve and AUC were used to evaluate different threshold criteria for model selection. To allow comparisons to be made across all of the approaches, AUC and ROC curves were calculated for all of the approaches.

Another goal of this chapter was to determine what conditions hinder or help approaches to correctly select the data generating model for a given observation. In Chapter 1, I hypothesized that

higher communality and more observation times would improve model selection. To assess these hypotheses, data were generated to have either low (.4) or high (.8) communality, and either 4 or 8 observation times. To evaluate the hypothesis that higher sample sizes would not improve model selection, data were generated to have either 200 or 1000 observations. To assess the hypothesis that the proportion of observations from given model does not impact model selection, data were generated such that the proportion of observations from the quadratic model was either 10%, 25%, or 40%. Lastly, to test the hypotheses that the greater the difference between the two models, the easier it would be for person level information to identify which model a given observation came from, data were generated to have either a smaller quadratic effect (1/20th the size of the linear effect) or a larger quadratic effect (1/10th the size of the linear effect).

Given the conditions that were used to test the hypotheses of interest, the simulation study in this chapter has 48 condition combinations. The simulation included three proportions of observations from the quadratic model (10%, 25%, or 40%). There were two sizes for the quadratic effect ($1/20^{th}$ or $1/10^{th}$ the size of the linear effect), two sample sizes (N=200, 1000), two number of observation times (4 or 8), and two levels of communality (.4 or .8). The next sections of Chapter 4 will describe the population models, data generation, fitted models, approaches for selecting a data generating model, methods for evaluating model selection and the results of this study.

Population Models

The population values for the covariance matrix and mean vector for the linear latent trajectory parameters were the same as those used in the first simulation study. The parameter estimates for the quadratic latent trajectory parameters were based on the same data used to obtain the linear latent trajectory parameters. McArdle and Bell (2000) did not fit a quadratic model to the data, but they did provide the means and covariance matrix for the data which I used to obtain parameter estimates for a quadratic LGCM. The estimated quadratic LGCM resulted in a quadratic effect that was approximately $1/18^{th}$ size of the linear effect, thus the parameter estimates for the small quadratic effect in this study

were based on slight modifications to the results from the quadratic LGCM fit to the observed means and covariance matrix for the data. The population covariance and mean structure for the small quadratic latent trajectories were:

$$\widehat{\boldsymbol{\Phi}}_{QS} = \begin{bmatrix} 80.75 & & \\ -.90 & 12.75 & \\ -.15 & -1.25 & 0.15 \end{bmatrix} \quad \widehat{\boldsymbol{\mu}}_{\eta_{QS}} = \begin{bmatrix} 30.0 \\ 10.0 \\ -0.5 \end{bmatrix}$$

The population model for the LGCM with a large quadratic effect was selected to preserve the correlations among the latent factors and the ratio of the quadratic mean and variance, and the ratio of the linear mean and variance, while changing the ratio of the quadratic effect to the linear effect from 1/20 to 1/10. This resulted in the following covariance matrix and mean vector:

$$\widehat{\boldsymbol{\Phi}}_{QL} = \begin{bmatrix} 80.75 & & \\ -1.18 & 19.13 & \\ -.26 & -2.64 & 0.45 \end{bmatrix} \quad \widehat{\boldsymbol{\mu}}_{\eta_{QL}} = \begin{bmatrix} 27.8 \\ 15.0 \\ -1.5 \end{bmatrix}$$

Together these matrices, along with the previously described matrices for the linear model, were used to generate the individual latent trajectory parameters from the linear and quadratic latent growth curve populations. Figure 31 is a plot of the mean trajectories for the linear (black line), small quadratic (blue line), and large quadratic (red line) models.

Data Generation

For each of the experimental conditions, 500 samples were generated to have either 200 or 1000 observations. Data from the quadratic population had either a small or large quadratic effect, as previously described, and were generated such that the proportion of observations from the quadratic model was either 10%, 25%, or 40%. After the individual latent trajectory parameters were generated using the appropriate population model, a linear or quadratic LGCM was used to transform the individual latent trajectory parameters into the observed data. For the conditions with 4 time points, the linear factor loading matrix from the previous studies were used and for the quadratic model the following factor loading matrix was used:

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$$

For the conditions with 8 time points, the linear factor loading matrix from the previous studies were used and for the quadratic model the following factor loading matrix was used:

$$\mathbf{\Lambda} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 12 \\ 1 & 5 & 25 \\ 1 & 6 & 36 \\ 1 & 7 & 49 \end{bmatrix}$$

All of the factor loading matrices specified that the intercept for the model was at time 1. For each observation, a time-specific residual (ε_{it}) was added at each time point such that the size of the variance of the time-specific residual ($VAR(\varepsilon_{it})$) depended on the target communality for a given data generating condition.

Fitted Models

Both a linear and a quadratic LGCM were fit to every data set from every data generating condition. The LGCMs were fit in SAS 9.2 using SAS PROC CALIS.

Selecting a Person-Level Data Generating Model

Four approaches for identifying the data generating model for person i were applied to the results from the fitted models. The four approaches were implemented using SAS PROC IML. Model selection was evaluated in several ways for each approach including true positive model selection, sensitivity, and specificity for the approaches with clear cut-off criteria, or a ROC curve and AUC for the conditions when clear cut-off criteria were not available. The specific criterion for each approach is described next.

For the first approach, examining factor score estimates by themselves, I examined factor score estimates from both the regression and Bartlett's methods. The factor score estimates were calculated

using Equation 5 and 6. For both methods, I tested if the estimate for the quadratic factor score estimate for a given observation was significantly different from zero. To do this, I divided an individual's quadratic factor score estimate by the quadratic factor score estimate's standard error and used a simple two-tailed z-test to calculate if the quadratic factor score estimate was significantly different from zero ($\alpha < .05$). To calculate the standard error for the z-test I used the following equations. For the regression factor score estimates (McDonald, 2011):

$$\mathrm{VAR}\left(\widehat{\boldsymbol{\eta}}_{rm}\right) = \left(\boldsymbol{\Lambda}'\Theta_{\varepsilon}^{-1}\boldsymbol{\Lambda} + \boldsymbol{\Phi}^{-1}\right)^{-1} \tag{18}$$

For Bartlett's factor score estimates (McDonald, 2011):

$$\mathrm{VAR}\left(\widehat{\boldsymbol{\eta}}_{GLS}\right) = \left(\boldsymbol{\Lambda}'\Theta_{\varepsilon}^{-1}\boldsymbol{\Lambda}\right)^{-1} \tag{19}$$

To calculate the standard error for a given individual's factor score estimate, I calculated the square root of the variance for the quadratic factor score estimates.

For the second approach, examining factor score residuals, I examined the difference between factor score residuals from the linear and quadratic latent curve models. A positive or near zero difference between the residuals indicated that there was little difference between the predictions for the linear and quadratic models, which suggested that the less complex linear model was the best approximation of the observed data. While a negative difference indicated that the predictions from the quadratic model were a better approximation of the observed data. I examined the difference in residuals using factor score residual analysis from both the regression and Bartlett's methods. To calculate the difference in residuals I used the Equation 8 to calculate the residuals for both the linear and quadratic models and then calculated the difference in the residuals for linear and quadratic models as:

$$\hat{\boldsymbol{\varepsilon}}_{i_{diff}} = \hat{\boldsymbol{\varepsilon}}_{i_{quadratic}} - \hat{\boldsymbol{\varepsilon}}_{i_{linear}} \tag{20}$$

which was equivalent to the difference in the predicted values:

$$\hat{\boldsymbol{\varepsilon}}_{i_{diff}} = \left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_{i_{quadratic}}\right) - \left(\boldsymbol{y}_i - \hat{\boldsymbol{y}}_{i_{linear}}\right) = \hat{\boldsymbol{y}}_{i_{linear}} - \hat{\boldsymbol{y}}_{i_{quadratic}} \quad (21)$$

The estimate for the variance matrix for the difference in residuals was calculated as:

$$VAR\left(\hat{\boldsymbol{\varepsilon}}_{i_{diff}}\right) = VAR\left(\hat{\boldsymbol{\varepsilon}}_{i_{quadratic}}\right) + VAR\left(\hat{\boldsymbol{\varepsilon}}_{i_{linear}}\right) - 2COV\left(\hat{\boldsymbol{\varepsilon}}_{i_{quadratic}}, \hat{\boldsymbol{\varepsilon}}_{i_{linear}}\right) \quad (22)$$

Because the covariance term in Equation 22 cannot be derived analytically, it was estimated empirically and for consistency the variance estimates for the quadratic and linear residuals were also empirically estimated. Some support for using the empirical estimates can be found Bollen and Arminger's (1991) work on the residual test statistic. They found that using the empirically estimated matrices to estimate the variances in Equations 10 and 11 resulted in similar conclusions to those found when using the variance estimates from the population matrices. To check that the empirically estimated variances would be similar to the variances estimated using Equations 10 or 11, I compared the estimates for several replications and found the estimates to be similar within 2 or 3 decimal places.

I substituted the values from Equation 20 and 22 for the residual and variance estimates in Equation 13 and used it to test if the difference in residuals is significantly different from zero. The test statistic for the difference in the linear and quadratic predicted values follows a chi-distribution with degrees of freedom equal to the number of time points and I used a cut-off criteria from the relevant chi-square distribution using $\alpha = .05$ as the significance level. In addition to this traditional cut-off criterion, I also examined a cut-off criterion that accounts for the issue of repeated significance testing by using a Bonferroni correction to the significance level.

This test of the difference in residuals was used as follows. First, if a given observation was found to have a difference in residuals that was not significantly different zero, the observation was classified as coming from the linear data generating model. If a given observation was found to have a difference in residuals that was significantly different from zero, then I classified the observation as from whichever model was a closer approximation of the observed data. If the quadratic model was a closer

approximation, then observation was classified as coming from the quadratic data generating model. If the linear model was a significantly closer approximation of the observed data, then observation was classified as coming from the linear data generating model.

For the third approach I examined the difference between the RMSR_i for the linear model and the quadratic model RMSR_i. The RMSR_i for both models was calculated using Equation 9 and then the difference was calculated by subtracting the RMSR_i for the linear model from the RMSR_i for the quadratic model as shown in the following equation:

$$\text{RMSR}_{i_{diff}} = \text{RMSR}_{i_{quadratic}} - \text{RMSR}_{i_{linear}} \qquad (23)$$

If the difference in RMSR_i resulted in positive or near zero values this indicated that the predicted values from the linear model were better or nearly the same as the predicted values from the quadratic model. Negative values of the difference in RMSR_i indicated that the quadratic model provided a better prediction of the observed data. Because the difference RMSR_i does not follow a known distribution, AUC and ROC curves were used to evaluate how different cut-off criteria impact the true positive and false positive rates.

For the fourth approach, I calculated a difference between the -2PLL_i for the linear model and the -2PLL_i for the quadratic model. The -2PLL_i for both models was calculated by taking two times Equation 14 and the difference was calculated by subtracting the -2PLL_i for the linear model from the -2PLL_i from the quadratic model. A negative value for the difference -2PLL_i indicated that the linear model was a better representation of the observed data, and positive values indicated that the quadratic model was a better representation of the observed data. Similar to the RMSR_i, the difference in -2PLL_i does not follow a known distribution and therefore was evaluated using AUC and ROC curves to assess how different cut-off criteria impact the true positive and false positive rates.

Evaluation of Model Selection Procedures

The performance of these approaches for identifying the data generating model was quantified in several ways. For the factor score estimate and factor score estimate residual approaches, I defined clear cut-off criteria to classify observations as coming from a linear or quadratic model. Therefore, to evaluate these approaches used true positive rate, sensitivity, and specificity. Similar to the study in Chapter 2, these calculations were made for each data set in the study. Within a given combination of experimental conditions, descriptive statistics, like mean, median, minimum, and maximum, were used to compare the approaches. For the approaches examining the difference in RMSR_i, and $\text{-2}PLL_i$, there was no clear cut-off criteria. Therefore, these approaches were evaluated using AUC and ROC curves to assess how different cut-off criteria impact the true positive rate (sensitivity) and false positive rate (specificity). AUC and ROC curves were calculated for each data set in the study. To compare the approaches with AUC, several descriptive statistics, like mean, median, minimum, and maximum, were examined. In addition, average ROC curves with observed 90% percentile intervals were also used to examine the performance of these approaches. To test the impact of levels within a condition, a series of ANOVA models were fit using the experimental conditions as predictors of AUC. The first ANOVA model tested the main effects of experimental condition, while the subsequent ANOVA models tested for interactions among the experimental conditions. ROC curves were then examined to visually examine how the experimental conditions impacted the performance of these approaches for classifying observations from the linear and quadratic models

These methods for quantifying the performance of these approaches allowed for comparisons between the factor score estimate and factor score estimate residual approaches, and between the RMSR_i and $\text{-2}PLL_i$ approaches. However, there was not a clear way to make comparisons across these pairs of approaches. Therefore to be able to compare across these approaches, AUC and ROC curves were also calculated for the factor score estimate and factor score estimate residual approaches, and the ANOVA models described for the RMSR_i and $\text{-2}PLL_i$ approaches, were estimated for the factor score estimate and

factor score estimate residual approaches. Together, the information gathered from this study provides evidence for what conditions improve or hinder the ability of these approaches to identify the data generating model for an individual and also allow comparisons among the approaches.

Results

Overall Performance

Table 22 contains the mean number of true positives within each combination of percent aberrant observations and sample size for the quadratic factor score estimates and difference in factor score residuals approaches. Table 22 shows that all of the approaches identified less than half of the observations from the quadratic model as being from the quadratic model. For the approaches examining the quadratic factor score estimates, Bartlett's factor score estimates identified almost twice as many observations as being from the quadratic model relative to the regression factor score estimates. The reverse pattern was found for the approaches examining the difference in the factor score residuals. These results seem to suggest that the Bartlett's factor score estimates more accurately estimate the true quadratic factor score estimate, but the factor score residuals from the regression method were better able to identify the data generating model relative to the Bartlett's factor score residuals. Table 22 also seems to indicate that the proportion of true positives appears to decrease slightly as the percent of aberrant observations increased.

As with the previous studies, sensitivity and specificity were also examined for the approaches examining the quadratic factor score estimates and difference in factor score residuals. Table 23 contains descriptive statistics for sensitivity across conditions. The mean, standard deviation, median, minimum, and maximum sensitivity for each approach confirms what was found in Table 22, which suggests that the approaches identified most of the observations from the quadratic model as being from the linear model. On average only 10 to 20 percent of the quadratic observations were identified as quadratic. The approaches examining the difference in factor score residuals did best overall. The difference in factor score residuals approach using regression factor score estimates had an average sensitivity of .39, while

the approach using Bartlett's factor score estimates had an average sensitivity of .27. Overall, the sensitivity results suggest that the cut-off criteria specified for these approaches wereAs not effective at correctly identifying the observations from the quadratic model.

Table 24 contains the results for specificity across conditions. The mean and median specificity suggest that all of the approaches correctly identified nearly all of the observations from the linear model as being from the linear model. The standard deviations suggest that with the exception of approaches examining the difference in factor score residuals, specificity was consistent across conditions. The approaches examining the difference in factor score residuals had higher standard deviations and showed a much larger difference between their minimum and maximum values, suggesting that these approaches did not have consistent specificity. The minimums specifically suggest that there were times when these approaches failed to identify half of the observations from the linear model as being from the linear model.

The results in Tables 22, 23 and 24 suggest that the approach evaluating whether the quadratic factor score estimate was significantly different from zero resulted in nearly all observations being classified as from a linear model, regardless of what model the observations were generated from. This provides some support for the hypothesis that examining factor score estimates directly may not always be useful for determining the model that was used to generate the data. In this case it could be that the difference between the two models was not great enough for the factor score estimates to clearly favor one model over the other, or it is possible that the cut-off criteria used to select the data generating model was not appropriate.

For the approaches examining the difference in factor score residuals Tables 22, 23 and 24 also indicate nearly all of the observations were classified as coming from the linear model regardless of the data generating model. This does not support the hypothesis that the approaches examining factor score residuals would do better than examining the factor score estimates directly. Together these results suggest that neither approach was able to successfully identify the correct data generating model given the conditions examined. Given how poorly these approaches performed, ANOVA models were not used to

examine the influence of the experimental conditions on the performance of these approaches using the specified cut-off criteria.

In addition to examining the true positives, sensitivity and specificity for the quadratic factor score estimates and difference in factor score residuals, I also examined AUC for these approaches as well as for the approaches using the difference in RMSR_i and $-2PLL_i$. Table 25 contains the descriptive statistic results for AUC. Table 25 indicates that on average the approaches performed similarly poor with an average AUC between .59 and .69. Table 25 shows that the approaches examining the quadratic factor score estimates had the highest average AUC followed by the Bartlett's factor score estimate residual and RMSR_i approaches. The $-2PLL_i$ approach had the lowest average AUC. These results do not support the hypothesis that the approaches examining the factor score estimates directly would do the worst and instead suggest that all the approaches perform about the same. The standard deviations in Table 25 ranged from .10 to .15 indicating that there was a lot of variability in AUC. The medians were lower than the means, indicating that across the conditions more than half of the AUCs were lower than the average AUC. The minimums for all of the approaches were less than .5, indicating that for at least one data set the approach did worse than chance classification. Figure 32 contains an example ROC plot from the results using the difference in regression factor score residuals when the AUC was .40. This example shows how the AUC can be less than .5 due to the approach either having high specificity with low sensitivity or low specificity with high sensitivity. An AUC less than .5 suggests that the approach was either classifying all the observations as from the linear model (high specificity and low sensitivity) or from the quadratic model (low specificity and high sensitivity). The maximum in Table 25 was 1 or near 1 for all of the approaches indicating that for at least one data set the approaches had perfect classification.

In addition to examining the descriptive statistic for AUC, average ROC curves for each approach were also examined. Figures 33 to 39 contain ROC curve plots of the mean ROC curve along with the 5^{th} and 95^{th} percentile for each of the approaches examined. The mean ROC curves indicate that the

approaches examining the quadratic factor score estimates were farthest from grey line representing chance classification, while the -2PLL_i approach had an average ROC curve that was closest to the chance classification line. The lines representing the 5th and 95th percentile for each approach suggest that the approaches varied a lot across the conditions examined. The 5th percentile for all of the approaches was close to the chance classification line, while the 95th percentile line to was greater or equal to .8 sensitivity for most levels of specificity indicating good classification. Together the results in Table 25 and Figures 33 to 39 indicate that there was a lot of variability in how well these approaches did at classifying observations from the linear and quadratic models, making it important to determine what factors influence the performance of these approaches.

Impact of Experimental Conditions

To test the hypotheses regarding the impact of communality, number of observations, sample size, proportion of quadratic model observations, and size of quadratic effect several ANOVA models were examined. For this study the ANOVA models were only used to determine the impact of the experimental conditions on AUC. The same analytic strategy from Chapter 2 was used to assess the impact of the experimental conditions on AUC in this chapter.

Table 26 contains the AUC ANOVA results for the approaches examining the quadratic factor score estimates. For both types of factor score estimates there was no meaningful main effect (η^2<.01) of sample size or percent of aberrant observations (in this case percent of observations from the quadratic model). These results support the hypotheses that the performance of these approaches would be unrelated to sample size or percent of aberrant observations. For the approach examining the quadratic factor score estimate using the regression factor score method there was a main effect of communality (η^2=.07), such that higher communality was associated with higher AUC. There were also meaningful main effects for the number of observation times (η^2=.49) and quadratic size (η^2=.24), as well as an interaction between these two factors (η^2=.09). Figure 40 shows the average ROC curves representing this interaction. It shows that more observation times and a larger quadratic are both associated with larger

AUC, but the effect of one is even greater at the higher level of the other factor. The combination of both 8 observation times and the larger quadratic effect resulted in an average ROC curve representing excellent classification (AUC>.90). These results support the hypotheses that more observation times and a larger difference between the two models would result in more accurate classification of observation.

For the approach examining the quadratic factor score estimates using Bartlett's factor score estimates there were also main effects for communality (η^2=.12), the number of observation times (η^2=.52) and quadratic size (η^2=.22w2), as well as an interaction between communality and number of observation times (η^2=.02) and also between number of observation times and quadratic size (η^2=.05). Figure 41 shows the average ROC curves representing the interaction between communality and number of observation times. It shows that higher communality and more observation times were associated with greater AUC and when both factors were at their highest level the average ROC curve showed good classification (AUC>.80). Figure 42 shows the average ROC curves representing the interaction between number of observation times and quadratic size. The results are similar to what was found for the regression quadratic factor score estimates approach such that higher levels of these factors were associated with higher AUC. Again at high levels of both factors the average ROC curve shows good classification. These results support the hypotheses that higher levels of communality, more observation times, and a greater quadratic effect would all be associated with better classification.

The AUC ANOVA results for the difference in factor score estimate residual and difference in RMSR_i approaches are show in Table 27. As with the approaches examining the quadratic factor score estimates directly there were no meaningful main effects for sample size (η^2<.01) or percent of aberrant observations (η^2<.01). There were meaningful main effects for number of observation times (η^2=.43-.53), communality (η^2=.08-.12) and quadratic size (η^2=.17-.25), as well as some interactions among these factors. For all of the approaches there was an interaction between communality and number of observation times (η^2=.02-.08), as well as an interaction between number of observation times and quadratic size (η^2=.08-.12). For the approach examining the difference in Barlett's RMSR_i there was also

an interaction between communality and quadratic size (η^2=.01). Figures 43 to 46 show the interactions between communality and number of observation times. They show the same pattern of results found for the previous interactions, with the approaches using regression factor score estimates showing better classification than the approaches using Bartlett's factor score estimates. Figures 47 to 50 show the interactions between number of observation times and quadratic size. Again they show that show the same pattern of results found for the previous interactions, although the classification appears to be better with more observation times and higher quadratic size, than when both number of observation times and communality are at the highest levels examined in this study. Figure 51 shows the interaction between communality and quadratic size for the difference in Barlett's RMSR_i approach. It shows a similar pattern to the previous interaction results, however the combination of high communality and greater quadratic effect size does not show as much improvement in the classification rate as the interactions with number of observation times show. Together, these results support the hypotheses that greater communality, number of observation times, and quadratic size result in better classification. They also suggest that for the approaches using regression factor score residuals when there is both a greater number of observation times and either higher communality or greater quadratic size these approaches do a good job of correctly classifying observations

Table 28 contains the AUC ANOVA results for the approach examining the difference in $-2PLL_i$. As with the previous approaches there was no meaningful effect of sample size (η^2<.01), supporting the hypothesis that sample size would be unrelated to the performance of this approach. There were meaningful main effects for number of observation times (η^2=.34), communality (η^2=.07), percent of aberrant observation (η^2=.08) and quadratic size (η^2=.14), as well as some interactions among these factors. As with the other approaches examined there was a meaningful interaction between communality and number of observation times (η^2=.03). There was also a meaningful three way interaction between number of observation times, quadratic size and percent of aberrant observations (η^2=.02). Figure 52 contains the interaction between communality and number of observation times. The results in Figure 52

suggest that as previously found higher communality and more observation times are associated with greater AUC, and that when both factors are high the approach does its best at correctly classifying observations. Figures 53 and 54 contain plots representing the interaction between number of observation times, quadratic size and percent of aberrant observations. Figure 53 shows mean ROC curves for all combinations of levels of number of observation times and percent of aberrant observations for the conditions with the small quadratic effect. While Figure 54 shows the same plots for the conditions with the large quadratic effect. The plots in Figure 53 and 54 show that more observation times and a larger quadratic effect are both associated with greater AUC, while a higher percent of aberrant observations (in this case, a higher number of quadratic observations) is associated with lower AUC. The plots show that the effect of percent of aberrant observations is largest when there are more observation times and a larger quadratic effect. In this case the classification goes from good (AUC=.82) with 10% of observations from the quadratic model to poor (AUC=.62) with 40% of observations from the quadratic model. These results suggests that the difference in -2PLL_i approach is best at classifying observations when communality, number of observation times, and the difference between the models of interest are larger and when most of the observations come from one of the models.

Overall the results from all of the AUC ANOVA models support the hypothesis that sample size would be unrelated to the ability of these approaches to correctly classify observations, while higher communality, number of observation times and difference in models were associated with better classification. Typically the approaches using regression factor score estimates performed better than using Bartlett's factor score estimates. One finding that was not expected was that for the approach examining the difference in -2PLL_i the percent of observations from a different model influenced the classification, such that the approach worked best when the majority of the observations were from one model. After examining the impact of the experimental factors on AUC it appears that all of these approaches can do an acceptable to good job of classifying observations, but there must be more

observation times (in this study 8 observation times was better than 4) and either greater communality or a greater difference between the models of interest.

Chapter 5

Discussion

Together these studies were designed to evaluate the performance of several approaches for using person level information to identify different types of aberrant observations in longitudinal data modeled by LGCMs. To achieve this goal three simulation studies were implemented each examining a different type of aberrance. The studies each assessed the overall performance of several factor score based approaches and person log-likelihood based approaches, under conditions that were hypothesized to enhance or impair the detection of aberrant observations. First I will discuss how the results of each study supported or failed to support my study specific hypotheses. Second I will discuss how the results of these studies fit with my overall hypotheses regarding these approaches to person level analysis. Finally, I will provide some recommendations for researchers based on the results of the three studies and end with a discussion of the limitations of these studies and some future directions for further research on person level analysis.

Discussion of Study Specific Results

Extreme Trajectory Aberrance Results

The goal of the first study was to examine if person level analysis could identify extreme trajectory aberrance. The results from this study showed that on average all of the person level analysis approaches were highly specific, but none of them were highly sensitive. The high specificity is at least partially due to the criteria selecting either 90% or 98% of the observations to be non-aberrant. If the non-aberrant observations were selected randomly, the specificity would be .81 or .96 by chance. Because the specificity results were greater than what would be expected by chance, there is some evidence that the approaches are able to correctly identify observations that do not have extreme trajectory aberrance.

However, the design of the study limits the conclusions that can be drawn regarding the specificity of these approaches.

Regarding the sensitivity of the approaches, the only approaches that were able to correctly identify more than half of the observations with extreme trajectory aberrance were the approaches examining the intercept factor score estimates directly (when only the intercepts were generated to be aberrant), and the approaches examining either factor score estimate directly and the -$2PLL_i$ approach (when both the intercepts and slopes were generated to be aberrant). These results fit with the hypotheses that the approaches examining factor score estimates directly and the -$2PLL_i$ approach would have better classification than the residual based approaches and IND_CHI_i approach when identifying extreme trajectory aberrance. However, the average sensitivity of even the best approaches was not very high, so it was useful to examine what factors were associated with sensitivity.

The results from the extreme trajectory aberrance study indicated that as hypothesized sample size was unrelated to sensitivity for the best approaches for detecting extreme trajectory aberrance. The -$2PLL_i$ approach did its best when both the intercepts and slopes were generated to be extreme and the data had high communality. This suggests that this approach was best at identifying individuals whose observations were closer to their underlying trajectory and who have trajectories that are very different from the average trajectory. The highest sensitivity for detecting extreme trajectory aberrance was found when using the approaches examining the factor score estimates directly. Sensitivity was greater than .8 for these approaches when there were either more observation times, or greater communality which supported my hypotheses regarding these factors. For both the approaches examining the factor score estimates directly and -$2PLL_i$ approach there was an effect of percent of aberrant observations such that more aberrant observations were associated with greater sensitivity.

Unlike the high specificity, the effect of percent of aberrant observations on sensitivity is probably not an artifact of the design of the study. If the 2% or 10% of the observations identified as aberrant were randomly selected, sensitivity would be .0004 or .01 by chance. The difference between

these chance levels of sensitivity is so small that is unlikely that it would explain very much of the variability in sensitivity. Therefore, one can interpret the effect of percent of aberrant observations as indicating that it is easier to detect observations with extreme trajectories when there are more aberrant observations. However, it is important to note that this effect is not likely to generalize such that as the percent of aberrant observations increases beyond the 10% in this study that sensitivity will continue to increase. First of all, as the proportion of observations with extreme trajectories becomes larger, at some point the variance estimates for the latent trajectory parameters would increase in response to the higher number of observations at the extremes. This will result in some of the observations with extreme trajectories no longer appearing extreme, given the variability in the trajectories expected based on the estimated model parameters. Secondly, if more than 10% of the observations in a given data set are considered to have extreme trajectories, one might wonder if the observations are truly extreme.

Regarding the use of regression versus Bartlett's factor score estimates there was not a clearly better approach. There was almost no difference between the regression and Bartlett's factor score estimates approaches that directly examined the intercepts. However, when examining both factor score estimates the regression approach did better, while the Bartlett's factor score approach did better when examining either factor score estimate. I had hypothesized that the number of observation times would have a greater impact on the approaches using regression factor score estimates. However, the difference in mean sensitivity for the high and low observation time groups was greater for the approaches using Bartlett's factor score estimates. In addition, number of observation times and communality tended to explain more of the variability in sensitivity for the Bartlett's factor score approach, suggesting that the performance of this approach may be more dependent on the levels of these factors.

Extreme Variability Aberrance Results

The second study evaluated whether person level analysis could identify extreme variability aberrance. The results from this study showed that on average all of the person level analysis approaches were highly specific, but only the factor score residual based approaches and -2PLL_i approach were

highly sensitive. These results fit with the hypotheses that the approaches examining factor score residual approaches and the -2PLL_i approach would do better than the approaches examining factor score estimates directly and the IND_CHI_i approach when identifying extreme variability aberrance. I had also hypothesized that the -2PLL_i approach would not do as well as the factor score residual based approaches, but the -2PLL_i approach was found to slightly outperform the factor score residual based approaches. This suggests that either comparing the observations to their model implied observations or to the overall model implied average observations will result in sufficient information with which to identify extreme variability aberrance.

Examination of the experimental factors impact on sensitivity resulted in some similar findings to those in the previous study. As with the first study, sample size was found to have no meaningful effect on the sensitivity of the best approaches for detecting extreme variability aberrance. This provides more evidence to support my hypothesis that sample size would be unrelated to the performance of these approaches. The -2PLL_i approach was found to be more sensitive with lower communality or more observation times, but the mean sensitivity was found to be over .9 for both of the levels of observation times and communality examined in this study. For the factor score residual based approaches there was no meaningful effect of communality, but a large effect size was found for observation times, such that more observation times were associated with greater sensitivity. As with the -2PLL_i approach, the sensitivity of the factor score residual based approaches tended to be high for both levels observation times. These results provided support for my hypothesis regarding number of observation times, but do not support my hypothesis regarding communality. Unlike the previous study, there was little or no effect of percent of aberrant for any of the best approaches for identify extreme variability aberrance. Relative to the previous study there was less variability in sensitivity to be explained by any of the experimental factors, which could be why the percent of aberrant observations was found to have no meaningful effect on sensitivity.

Regarding the use of regression versus Bartlett's scores, the approaches using regression scores were more sensitivity. However, there was little difference for the $RMSR_i$ approach. The type of factor score estimate mattered more for factor score estimate residual test statistic, where the regression factor score approach had an average sensitivity of .93 versus the Bartlett's approach average sensitivity of .78. These results may be due to the shrinkage of the regression scores producing residuals for the aberrant observations that were larger than the residuals from the Bartlett's approach, making it easier to identify observations with extreme variability aberrance. Similar to the previous study and counter to my hypothesized relationship, number of observation times tended to have a greater impact on the approaches using Bartlett's scores.

Extreme Functional Form Aberrance Results

The third study examined whether person level analysis could be used to identify functional form aberrance. This study focused on differentiating observations generated from a quadratic model, from those generated from a linear model. Unlike the previous studies, cut-off criteria were only examined for a subset of the approaches, specifically the approaches examining the quadratic factor score estimates directly and the approach examining the difference in factor score residuals. The results using cut-off criteria found that most of the observations were identified as being linear, regardless of their data generating model, suggesting that the specified criteria were not effective at detecting functional form aberrance.

In addition to examining cut-off criteria for a subset of the approaches, AUC and ROC curves were examined for all of the approaches. The AUC results indicated that approaches examining the quadratic factor score estimates had the highest average AUC followed by the Bartlett's factor score estimate residual and RMSR approaches. These results did not support my hypothesis that the approaches examining the factor score estimates directly would do the worst and instead suggest that all the approaches perform about the same on average.

Even though the average AUC results indicated poor classification (AUC <.80) for all of the approaches, the 5th and 90th percentiles around the mean ROC curves as well as the standard deviations for AUC indicated that there was a lot of variability in the classification of these approaches. It was therefore useful to examine which of the experimental factors could explain this variability. With the exception of the -2PLL_i approach, the results from the AUC ANOVAs supported the hypotheses that sample size and percent of aberrant observations were not associated with better classification. They also supported the hypotheses that higher communality, more observation times, and a larger difference between the models would all be associated with better identification of functional form aberrance.

Examination of the experimental factors identified conditions with good average classification (AUC >.80). The interaction results from the AUC ANOVAs indicated that with more observation times and either higher communality or a larger difference between the data generating models the average AUC was at or above .80 for nearly all of the approaches. These results suggest that any of these approaches could be used to identify functional form aberrance, but only under specific conditions. These findings were true for both the regression and Bartlett's factor score methods, although the regression factor score estimates tended to have slightly higher AUCs than Bartlett's method. Unlike the previous studies, the results showed a larger difference in mean sensitivity for the high and low observation time groups for the approaches using regression factor score estimates and the number of observation times tended to explain more of the variability in sensitivity for the approaches using regression factor score estimates.

As previously stated the only approach which performed poorly at identifying functional form aberrance across nearly all of the conditions was the -2PLL_i approach. The results for this approach suggested that it was influenced by all of the factors except sample size, so it is possible that in situations with higher communality, more observation times, or a greater difference between the two models of interest this approach would show better classification. However, under the conditions examined in this study, this approach was the least promising across nearly all the levels of these factors. In fact, the only

time the -2PLL_i approach was found to have good classification (AUC=.82) when there were more observation times, a low proportion of observations from the quadratic model and a larger difference between the linear and quadratic data generating models. One result that was unique to this approach and which should be considered in future studies of this approach was that the percent of aberrant observations explained a meaningful amount of variability in AUC. The mean AUC for each level of this factor indicates that as the proportion of observations from the aberrant model (the quadratic model) increased, the AUC decreased. This suggests that the ability of the -2PLL_i approach to classify observations from different models, depends in part on how many observations in the data are from each model. Because the estimated model parameters are used as part of the calculation for the -2PLL_i approach, this result may be due to the parameter estimates for each of the fitted models being further from the population parameters, when the data are more equally split between the two models of interest.

Discussion of Overall Performance of Person Level Analysis

Overall, the results of these three studies suggest that different person level analysis approaches worked best for different types of aberrance. Examining factor score estimates directly did the best at identifying extreme trajectory aberrance, whereas the approaches using factor score residuals and the -2PLL_i approach were the best approaches for identifying extreme variability aberrance. The AUC and ROC curve results from the third study suggest that under certain conditions, either examining factor score estimates directly or the factor score residual based approaches can provide some information regarding functional form aberrance. These results suggest that if a researcher is concerned about a specific type of aberrance, they should examine the person level fit index that captures the information that is relevant to the aberrance they are concerned about. Alternatively, if a researcher does not have a specific form of aberrance in mind, but is just generally concerned about aberrant observations influencing their results, the researcher should examine several person fit indices.

In addition to identifying promising approaches for identifying different types of aberrant observations, the results of these studies also provided information regarding the three main hypotheses I had about what factors would be associated with the performance of these person level approaches. First, I hypothesized that the approaches would do better when data were closer to the data generating model. To assess this hypothesis for each type of aberrance I examined two levels of communality. I found that for the approaches that worked best at identifying a given type of aberrant observations higher communality was associated with better identification of aberrant observations, but only for extreme trajectory aberrance or functional form aberrance. For the approaches best at detecting extreme variability aberrance there was either no meaningful effect of communality or for the -2PLL_i approach higher communality was associated with slightly worse sensitivity. This suggests that communality is less important when trying to assess extreme variability. This result did not fit with my expectation that lower communality would make it more difficult to identify extreme variability aberrance. However, it is possible that this result is an artifact of the size of the variability chosen to generate the aberrant observations, such that the variability of the aberrant observations is so large relative to the non-aberrant observation that the communality did not impact the sensitivity of these approaches.

My second hypothesis was that more observation times would be associated with a greater ability for the person level approaches to be able to identify aberrant observations. I found that across the three types of aberrance examined in these studies, number of observation times was consistently shown to be associated with a greater ability to correctly classify observations as aberrant or non-aberrant. This suggests that researchers can be more confident that they have correctly identified aberrant observations when they have more observations for a given individual. Related to this hypothesis, I also thought that number of observation times would be more important for the approaches using regression factor score estimates. I found this to be supported when using the approaches to identify functional form aberrance, but the reverse was found when trying to identify extreme trajectory and extreme variability aberrance. It

is possible that the shrinkage effect on the regression scores was not as strongly associated with observation times as I had hypothesized.

Lastly, my third hypothesis was that sample size would not influence the ability of the person level approaches to identify aberrant observations. For all the types of aberrance examined, sample size did not explain a meaningful amount of variance for any of the approaches which were better at identifying a given type of aberrance. This provides some support for my hypothesis. However, it is possible that this is a result of having a correctly specified model for the majority of the observations in any given data set, such that the model parameter estimates were able to produce accurate individual fit measures. For example, in a situation where sample size is strongly related to the accuracy of the model parameter estimates it may be possible that sample size would influence the ability of the person level approaches to identify aberrant observations.

Together these findings suggest that researchers can use person level information to assess whether aberrant observations are in their data. These approaches will work better with more observation times and greater communality. These factors are especially important when assess functional form aberrance, along with the difference between the two models being compared. These findings should encourage researchers to examine person level information when assessing the fit of a model or when comparing models.

Limitations of the Current Research

As with all studies, the results of this research have limitations, the first being that these results have occurred while studying a simple LGCM. A simple LGCM was chosen, because it was not clear whether any of these person level approaches would work and I believed the best place to start to assess this would be in a 'best case scenario' where the LGCM did not have any predictors and was correctly specified. This resulted in data generating conditions in which the overall model fit for the LGCMs was

very good and it is possible that with a misspecified or otherwise poorly fitting model these results may not apply.

Besides examining a simple LGCM, these studies were also limited to the types of aberrance examined. One type of aberrance that was not examined was time-specific aberrance. The results of these studies suggest that the approaches examining the factor score estimates directly would only be able to detect this type of aberrance if the time-specific aberrance resulted in one or more of the factor score estimates for a given individual to appear to be extreme. In order for the factor score residual based approaches to be able to identify time-specific aberrance, the aberrant observation time would need to occur, such that the factor score estimates imply an underlying trajectory that does not represent any of the observed data for the individual. This would result in larger residuals for that individual's observations and therefore larger estimates from the residual-based indices. Lastly, given the findings of these studies it is unlikely that the -2PLL_i approach would be able to detect time-specific aberrance. When examining extreme trajectory aberrance the -2PLL_i approach did better when both the intercept and slope were extreme, which means that the time-specific aberrance would need to occur such that the entire model implied trajectory for a given individual would appear sufficiently different from the average trajectory.

Another type of aberrance that was not examined by these studies is unbalanced extreme trajectory aberrance. The first study examined extreme trajectory aberrance with an equal number of extremely low and extremely high trajectories. It is possible that there would be situations where there might be a cluster of extreme trajectories that are only lower or higher than average. An example would be a study of anxiety over time in a non-clinical sample. Most individuals in this sort of sample will experience a few symptoms of anxiety while a smaller group of individuals may experience many anxiety symptoms. This would create a skewed distribution of observations, unlike the data examined in the first study. This skewed data could result in model parameter estimates that are pulled towards the cluster of aberrant observations. These inaccurate model parameter estimates could in turn result in inaccurate

factor score estimates, which may make the aberrant observations appear less aberrant and therefore harder to identify.

Besides being limited to a few types of aberrant observations, the first two studies and part of the third study focused on cut-off criteria for classifying observations as aberrant or non-aberrant. In practice researchers do not know how many aberrant observations should be in the data and therefore do not know what the optimal cut-off criteria should be. In these studies, a limitation of the cut-off criteria was that it is at least partially responsible for the high specificity that was observed, because the majority of the observations would be identified as non-aberrant. More generally the use of cut-off criteria can serve to mislead researchers into thinking they have done a thorough investigation of their data, when it is possible that they have included too many or too few observations as aberrant. In practice researchers may want to us cut-off criteria to identify observations worth further visual investigation. Alternatively, researchers may want to examine the distribution of person fit information with histograms, box plots or scatterplots, in order to identify aberrant observations.

A limitation of the study on functional form aberrance is that it is possible that some of the observations that were from a linear or quadratic model could have been identified as being from the other model for reasons other than the person level analysis approach being incorrect. For example, an observation could be incorrectly identified simply because the added time-specific variability resulted in the individual's observations taking on the appearance of the other model. Alternatively, the size of the quadratic effect could have been so small that there was little practical difference between the linear and quadratic models. Both of these possibilities, may explain why the person level approaches for comparing model fit at the person level needed more time points and either higher communality or a greater difference between the two models in order to have an average AUC that indicated good classification.

Another potential limitation of the study on functional form aberrance is that it implies that every individual has a true data generating functional form. With observed data we may believe that there isn't a

true underlying functional form, but that the LGCMs serve as a way to approximate the relationships among the observed data. The design of the study on functional form aberrance assessed whether the correct data generating model was selected, but if we don't believe that either model is correct, such information may not seem useful. Nevertheless, the results of the third study do provide some information regarding each model as an approximation of the observed data. The results of the third study suggest that person level information can be used to make a decision regarding whether the approximations from the two models of interest are significantly different from each other, and then to decide which model is closer to the observed data for a given individual. This may be useful in situations where researchers are trying to decide whether a more complex model is better than a simpler model and the overall model fit indices and/or the parameter estimate significance tests are giving conflicting information. For example, it is possible that one could find that the overall model fit of the quadratic model is better that the linear model while the estimates associated with the quadratic factor are near zero suggesting that on average the quadratic effect is not meaningful. Examining the person level information could provide more information on whether any individuals appear to have a meaningful nonlinear pattern to their observed data. Therefore, even if researchers believe that both models are just approximations, the person level information may help the researcher to select a better approximation.

Recommendations for Researchers

Given the results and the limitations of these studies, there are a few recommendations that can be made for researchers wondering how to identify aberrant observations in their data. First, there is evidence from these simulation studies that different approaches work well for different types of aberrance. The results from the first simulation study suggest that examining factor score estimates can help to identify individuals with extreme trajectories, but researchers will need to either have higher communality or more observation times to be confident that the factor score estimates can accurately identify individuals with extreme trajectory aberrance. The results of the second study suggest that the -$2PLL_i$ approach or the factor score estimate residual test using regression factor score estimates or the

$RMSR_i$ approaches using either type of factor score estimates are all able to consistently identify extreme variability aberrance across all the levels of the conditions examined. Finally, the third study suggests that examining the quadratic factor score estimates, or the difference in the factor score estimate residual statistic, or the difference in $RMSR_i$ can help to compare the fit of two models for a given individual. However, like the extreme trajectory aberrance, these approaches are more accurate when there are more observation times and either higher communality or a greater difference between the models being examined.

These studies examined several cut-off criteria, including using the proportion of aberrant observations used to generate the data. This was done to examine ideal conditions for detecting aberrant observations, which may not be as useful in practical applications unless researchers have a hypothesis regarding the proportion of observations they expect to be aberrant. Even though researchers will not know the optimal cut-off criteria to apply, the results of these studies can provide some information regarding cut-off criteria. One finding related to the cut-off criteria that was consistent across the studies, was that the Bonferroni correction for the factor score residual analysis statistic, was overly conservative and therefore not recommended. In addition, even though the AUC and ROC curve results from the third suggest that it is useful to examine the size of the quadratic factor score estimate when comparing a quadratic and a linear model. The results from the cut-off criteria in the third study suggest that researchers should not simply test a confidence interval around the quadratic factor score estimate. Furthermore, the results from the third study suggest that researchers can use criteria to accurately classify observations generated from different models. However, the results do not suggest a specific cut-off criteria to use. In practice researchers will need to determine appropriate cut-off criteria by considering the trade-off between sensitivity, identifying observations from a more complex model, versus specificity, identifying observations from a more parsimonious model, when using the proposed approaches. For all the types of aberrance, researchers should consider the costs of misclassifying observations and set any criteria for classification accordingly.

Lastly, in these studies the number and type of aberrant observations were known. In practice researchers may not know how many or what type of aberrant observations are in their data. Because the number of aberrant observations is unknown, researchers are encouraged to examine the distributions of person level fit information to see if any patterns occur, rather than relying on cut-off criteria alone. By examining the distributions of person level fit indices researchers may be able to identify a group of observations which appear to be distinct from the majority of the observations and therefore worthy of further investigation. In addition to examining the distribution of the person level fit information, researchers are also encouraged to examine the observed data for the observations that have person level fit indices that suggest aberrance. One way to examine the observed data would be to plot the observed data and the model-implied means, in this case the model implied average trajectory. Visually examining the observed data, especially relative to the model-implied pattern of observations, should provide some information about the type of aberrance that the person level fit indices have identified. It may also reveal data entry errors or individuals who are worth further investigation regarding their difference from the model-implied average observation. This combination of using statistical information to inform which observations are worth visual examination should provide the most efficient means for identifying meaningfully aberrant observations.

Future Directions

In order to better understand how effective these approaches are for assessing different types of aberrance future studies should focus on examining different types of aberrance and applications of these approaches in more complex models. Some other types of aberrance that would be useful to examine are time-specific aberrance and unbalanced extreme trajectory aberrance. Studying time-specific aberrance would be useful for any type of data, while studying unbalanced extreme trajectory aberrance will be especially useful for constructs that typically result in observed data with a skewed distribution.

The models in this study were very simple and in practice researchers will need to know how well these techniques work in more complex models. It would be useful to examine models with covariates predicting the functional form. Using covariates should result in better factor score estimates, especially for the approaches using the regression method, which could result in more accurate assessments of person level fit. However, the improvement in the factor score estimates will probably be related to the reliability of the covariates and the strength of their relationship with the latent factors, such that more reliable covariates with stronger relationships to the latent factors will have a greater impact on the factor score estimates and subsequent person fit indices.

Given that researchers never know for sure if they have specified the LGCM correctly it would be especially useful to apply these person level fit approaches in situations where the model has been misspecified in some way. It is possible that the person level fit approaches will be able to detect some misspecifications, while other misspecifications may result in person level indices that are not informative. For example, misspecifying the time specific error structure might result in higher person level residual statistics. Upon examination of the individuals with high person level residual statistics, a researcher may find the error in the model specification. Alternatively, if a researcher has misspecified a model by not including an important covariate predicting the latent trajectory factors an examination of the factor score estimates may result in some individuals appearing to have extreme trajectories. However, had the covariate been included in the model their trajectory would be considered normal.

These are just a few areas for future research into person level analysis. Given the current research, there is evidence that it is worthwhile to examine person level information. Future research should help to provide more recommendations for how to best apply these approaches in practice.

Table 1. Mean number of true positives across conditions for Study 1 examining extreme intercept trajectory aberrance.

		Percent Aberrant Observations			
		2%		10%	
Method	Factor Score Type	N=200, Aber=4	N=1000, Aber=20	N=200, Aber=20	N=1000, Aber=100
Intercept FS	Regression	2.82	13.83	16.58	82.61
	Bartlett's	2.82	13.76	16.30	81.28
FS Residuals	Regression	.21	1.10	2.87	14.48
	Bartlett's	.02	.09	.55	2.93
Bonf. FS Residuals	Regression	.00	.00	.04	.05
	Bartlett's	.00	.00	.00	.00
RMSR_i	Regression	.13	.60	2.37	11.62
	Bartlett's	.08	.38	2.06	10.00
$-2PLL_i$	-	1.06	5.16	7.40	36.80
IND_CHI_i	-	.28	1.57	3.47	18.96

Notes. Sim N = 2000; N= the sample size in each simulated data set; Aber = the total number of aberrant observations in each simulated data set;

Table 2. Mean number of true positives across conditions for Study 1 examining extreme intercept and slope trajectory aberrance.

		Percent Aberrant Observations			
		2%		10%	
Method	Factor Score Type	N=200, Aber=4	N=1000, Aber=20	N=200, Aber=20	N=1000, Aber=100
Both FS	Regression	1.63	8.67	9.84	50.54
	Bartlett's	.87	4.34	7.74	38.78
Either FS	Regression	2.40	11.89	14.85	74.09
	Bartlett's	2.67	13.28	16.36	81.82
FS Residuals	Regression	.49	2.48	4.36	22.16
	Bartlett's	.02	.09	.54	2.91
Bonf. FS Residuals	Regression	.01	.03	.08	.14
	Bartlett's	.00	.00	.00	.00
RMSR_i	Regression	.30	1.53	3.62	17.88
	Bartlett's	.08	.41	1.99	10.02
$-2PLL_i$	-	1.99	9.82	11.07	55.16
IND_CHI_i	-	.30	1.83	3.57	20.22

Notes. Sim N = 2000; N= the sample size in each simulated data set; Aber = the total number of aberrant observations in each simulated data set; Both FS = both factor score estimates were identified as aberrant; Either FS = either the intercept or the slope factor score estimate was identified as aberrant;

Table 3. Sensitivity across conditions for Study 1 examining extreme intercept trajectory aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Intercept FS	Regression	.76	.17	.80	.00	1.00
	Bartlett's	.76	.21	.85	.00	1.00
FS Residuals	Regression	.10	.09	.10	.00	.50
	Bartlett's	.02	.03	.00	.00	.25
Bonf. FS Residuals	Regression	.00	.01	.00	.00	.25
	Bartlett's	.00	.00	.00	.00	.05
RMSR_i	Regression	.07	.08	.05	.00	.75
	Bartlett's	.06	.07	.05	.00	.50
$-2PLL_i$	-	.32	.14	.31	.00	1.00
IND_CHI_i	-	.13	.10	.13	.00	.75

Note. Sim N = 8000

Table 4. Sensitivity across conditions for Study 1 examining extreme intercept and slope trajectory aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Both FS	Regression	.46	.14	.50	.00	1.00
	Bartlett's	.30	.24	.25	.00	1.00
Either FS	Regression	.67	.16	.65	.00	1.00
	Bartlett's	.74	.20	.75	.00	1.00
FS Residuals	Regression	.17	.12	.19	.00	.75
	Bartlett's	.02	.03	.00	.00	.25
Bonf. FS Residuals	Regression	.00	.02	.00	.00	.25
	Bartlett's	.00	.00	.00	.00	.05
RMSR_i	Regression	.13	.10	.14	.00	.75
	Bartlett's	.06	.07	.05	.00	.50
$-2PLL_i$	-	.52	.15	.50	.00	1.00
IND_CHI_i	-	.14	.10	.15	.00	.75

Notes. Sim N = 8000; Both FS = both factor score estimates were identified as aberrant; Either FS = either the intercept or the slope factor score estimate was identified as aberrant;

Table 5. Specificity across conditions for Study 1 examining extreme intercept trajectory aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Intercept FS	Regression	.99	.01	.99	.93	1.00
	Bartlett's	.99	.01	.99	.92	1.00
FS Residuals	Regression	.93	.04	.94	.83	1.00
	Bartlett's	.98	.02	.99	.93	1.00
Bonf. FS Residuals	Regression	1.00	.00	1.00	.98	1.00
	Bartlett's	1.00	.00	1.00	.99	1.00
RMSR_i	Regression	.94	.04	.96	.89	1.00
	Bartlett's	.94	.04	.95	.89	.99
$\text{-2}PLL_i$	-	.96	.03	.97	.89	1.00
IND_CHI_i	-	.95	.04	.96	.89	1.00

Note. Sim N = 8000

Table 6. Specificity across conditions for Study 1 examining extreme intercept and slope trajectory aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Both FS	Regression	1.00	.01	1.00	.94	1.00
	Bartlett's	1.00	.01	1.00	.96	1.00
Either FS	Regression	.95	.03	.97	.86	1.00
	Bartlett's	.95	.04	.97	.84	1.00
FS Residuals	Regression	.94	.04	.94	.84	1.00
	Bartlett's	.98	.02	.99	.92	1.00
Bonf. FS Residuals	Regression	1.00	.00	1.00	.98	1.00
	Bartlett's	1.00	.00	1.00	.99	1.00
RMSR_i	Regression	.95	.04	.96	.89	1.00
	Bartlett's	.94	.04	.95	.89	.99
$-2PLL_i$	-	.97	.02	.98	.91	1.00
IND_CHI_i	-	.95	.04	.97	.89	1.00

Notes. Sim N = 8000; Both FS = both factor score estimates were identified as aberrant; Either FS = either the intercept or the slope factor score estimate was identified as aberrant;

Table 7. Sensitivity means by experimental condition and final ANOVA model results for methods examining factor score estimates and factor score residual approaches for Study 1 examining extreme intercept trajectory aberrance.

		Factor Score Type									
		Regression					Bartlett's				
Method	Predictor	M - Low*	M - High*	DF	F	η^2	M - Low*	M - High*	DF	F	η^2
Intercept FS	Sample size	.77	.76	1	9.28	.00	.76	.75	1	13.67	.00
	Obs. times	.70	.83	1	2546.75	.15	.64	.87	1	8180.23	.27
	Communality	.71	.82	1	1934.68	.11	.66	.85	1	5473.46	.18
	% Aberrant	.70	.83	1	2466.34	.14	.70	.81	1	2289.51	.08
	Obs. times* Communality	-	-	1	2095.58	.12	-	-	1	5561.15	.19
	Obs. times*%Aberrant	-	-	-	-	-	-	-	1	319.72	.01
FS Residuals	Sample size	.10	.10	1	.95	.00	.02	.02	1	.56	.00
	Obs. times	.11	.09	1	46.36	.00	.01	.02	1	260.22	.03
	Communality	.10	.10	1	.31	.00	.02	.02	1	.32	.00
	% Aberrant	.05	.14	1	2996.99	.27	.01	.03	1	1471.41	.15
	Obs. times*%Aberrant	-	-	-	-	-	-	-	1	132.85	.01
$RMSR_i$	Sample size	.08	.07	1	2.84	.00	.06	.06	1	3.84	.00
	Obs. times	.08	.07	1	41.32	.00	.06	.06	1	.32	.00
	Communality	.08	.07	1	14.92	.00	.06	.06	1	.15	.00
	% Aberrant	.03	.12	1	4049.08	.33	.02	.10	1	4796.32	.37

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 8. Sensitivity means by experimental condition and final ANOVA model results for methods examining factor score estimates for Study 1 examining extreme intercept and slope trajectory aberrance.

		Factor Score Type									
		Regression					Bartlett's				
Method	Predictor	M - Low*	M - High*	DF	F	η^2	M - Low*	M - High*	DF	F	η^2
Both FS	Sample size	.45	.47	1	70.09	.01	.30	.30	1	.00	.00
	Obs. times	.40	.52	1	2494.85	.18	.19	.42	1	9252.17	.23
	Communality	.41	.51	1	2033.25	.14	.14	.46	1	17509.60	.43
	% Aberrant	.42	.50	1	1085.59	.08	.22	.39	1	4948.14	.12
	Obs. times* Communality	-	-	1	29.63	.00	-	-	1	1199.21	.03
	Obs. times * %Aberrant	-	-	1	16.33	.00	-	-	-	-	-
	Communality* %Aberrant	-	-	1	52.12	.00	-	-	-	-	-
	Obs. times * Communality* %Aberrant	-	-	1	320.92	.02	-	-	-	-	-
Either FS	Sample size	.67	.67	1	3.24	.00	.74	.74	1	.70	.00
	Obs. times	.64	.70	1	1277.21	.05	.68	.81	1	3095.17	.11
	Communality	.58	.76	1	9404.61	.37	.61	.88	1	12477.20	.44
	% Aberrant	.60	.74	1	5525.54	.21	.67	.82	1	4016.48	.14
	Obs. times* Communality	-	-	1	1520.49	.06	-	-	-	-	-
	Communality* %Aberrant	-	-	-	-	-	-	-	1	864.03	.03

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 9. Sensitivity means by experimental condition and final ANOVA model results for factor score estimate residual methods for Study 1 examining extreme intercept and slope trajectory aberrance.

		Factor Score Type									
		Regression					Bartlett's				
Method	Predictor	M - Low*	M - High*	DF	F	η^2	M - Low*	M - High*	DF	F	η^2
FS Residuals	Sample size	.17	.17	1	1.00	.00	.02	.02	1	3.32	.00
	Obs. times	.19	.15	1	379.66	.03	.01	.02	1	247.35	.02
	Communality	.15	.19	1	378.16	.03	.02	.02	1	.04	.00
	% Aberrant	.12	.22	1	1887.53	.17	.00	.03	1	1616.66	.16
	Obs. times* Communality	-	-	1	273.84	.03	-	-	-	-	-
	Obs. times*%Aberrant	-	-	-	-	-	-	-	1	123.41	.01
$RMSR_i$	Sample size	.13	.13	1	.02	.00	.06	.06	1	.34	.00
	Obs. times	.15	.11	1	439.65	.04	.06	.06	1	.25	.00
	Communality	.12	.13	1	14.25	.00	.06	.06	1	.07	.00
	% Aberrant	.08	.18	1	3194.61	.27	.02	.10	1	4761.07	.37
	Obs. times* Communality	-	-	1	170.26	.01	-	-	-	-	-

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 10. Sensitivity means by experimental condition and final ANOVA model results for log-likelihood methods for Study 1 examining extreme intercept trajectory aberrance.

Method	Predictor	M - Low*	M - High*	DF	F	η^2
-2PLL_i	Sample size	.32	.31	1	3.55	.00
	Obs. times	.35	.29	1	535.20	.05
	Communality	.29	.34	1	411.24	.04
	% Aberrant	.26	.37	1	1687.67	.15
	Obs. times* Communality	-	-	1	463.64	.04
IND_CHI_i	Sample size	.12	.13	1	44.89	.00
	Obs. times	.15	.11	1	510.78	.04
	Communality	.12	.13	1	49.32	.00
	% Aberrant	.07	.18	1	3137.27	.27

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 11. Sensitivity means by experimental condition and final ANOVA model results for log-likelihood methods for Study 1 examining extreme intercept and slope trajectory aberrance.

Method	Predictor	M - Low*	M - High*	DF	F	η^2
-2PLL_i	Sample size	.53	.52	1	3.60	.00
	Obs. times	.52	.53	1	4.44	.00
	Communality	.42	.63	1	7167.04	.46
	% Aberrant	.49	.55	1	566.39	.04
IND_CHI_i	Sample size	.13	.15	1	110.14	.01
	Obs. times	.16	.12	1	357.47	.03
	Communality	.12	.15	1	189.64	.02
	% Aberrant	.08	.19	1	3116.44	.26

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 12. Specificity means by experimental condition and final ANOVA model results for methods examining factor score estimates and factor score residual approaches for Study 1 examining extreme intercept trajectory aberrance.

		Factor Score Type									
		Regression					Bartlett's				
Method	Predictor	M - Low*	M - High*	DF	F	η^2	M - Low*	M - High*	DF	F	η^2
Intercept FS	Sample size	.99	.99	1	5.63	.00	.99	.99	1	5.31	.00
	Obs. times	.98	.99	1	5070.84	.12	.98	.99	1	15989.40	.20
	Communality	.98	.99	1	4002.68	.10	.98	.99	1	11129.50	.14
	% Aberrant	.99	.98	1	16813.00	.40	.99	.98	1	21553.00	.27
	Obs. times* Communality	-	-	1	4025.15	.10	-	-	1	10965.90	.14
	Obs. times * %Aberrant	-	-	1	1549.48	.04	-	-	1	5303.45	.07
	Communality* %Aberrant	-	-	1	1305.44	.03	-	-	1	3910.84	.05
	Obs. times * Communality* %Aberrant	-	-	1	1164.21	.03	-	-	1	3686.20	.05
FS Residuals	Sample size	.93	.93	1	40.36	.00	.99	.98	1	71.59	.00
	Obs. times	.93	.93	1	20.19	.00	.99	.98	1	4754.56	.10
	Communality	.93	.93	1	234.45	.00	.98	.98	1	.02	.00
	% Aberrant	.97	.89	1	114741.00	.93	1.00	.97	1	30450.30	.66
	Obs. times*%Aberrant	-	-	-	-	-	-	-	1	2525.99	.06
$RMSR_i$	Sample size	.94	.94	1	3.46	.00	.94	.94	1	3.18	.00
	Obs. times	.94	.94	1	64.97	.00	.94	.94	1	1.90	.00
	Communality	.94	.94	1	27.02	.00	.94	.94	1	.44	.00
	% Aberrant	.98	.90	1	651813.00	.99	.98	.90	1	775201.00	.99

*Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 13. Specificity means by experimental condition and final ANOVA model results for methods examining factor score estimates for Study 1 examining extreme intercept and slope trajectory aberrance.

		Factor Score Type									
		Regression					Bartlett's				
Method	Predictor	M - Low*	M - High*	DF	F	η^2	M - Low*	M - High*	DF	F	η^2
Both FS	Sample size	1.00	1.00	1	.55	.00	1.00	1.00	1	2.35	.00
	Obs. times	.99	1.00	1	748.10	.05	1.00	1.00	1	1824.24	.07
	Communality	.99	1.00	1	1892.42	.14	.99	1.00	1	6318.51	.23
	% Aberrant	1.00	.99	1	1940.20	.14	1.00	1.00	1	5163.63	.19
	Obs. times* Communality	-	-	1	370.44	.03	-	-	1	915.86	.03
	Obs. times * %Aberrant	-	-	1	867.43	.06	-	-	1	992.06	.04
	Communality* %Aberrant	-	-	-	-	-	-	-	1	3394.59	.13
	Obs. times * Communality* %Aberrant	-	-	-	-	-	-	-	1	461.61	.02
Either FS	Sample size	.95	.95	1	15.25	.00	.95	.95	1	1.52	.00
	Obs. times	.95	.96	1	1740.78	.02	.94	.96	1	14182.40	.07
	Communality	.94	.96	1	4866.27	.06	.93	.96	1	31944.40	.17
	% Aberrant	.98	.92	1	66208.90	.76	.98	.92	1	119524.00	.62
	Obs. times* Communality	-	-	1	1519.75	.02	-	-	-	-	-
	Obs. times * %Aberrant	-	-	1	905.54	.01	-	-	1	6284.09	.03
	Communality* %Aberrant	-	-	1	2998.84	.03	-	-	1	12100.30	.06
	Obs. times * Communality* %Aberrant	-	-	1	1205.72	.01	-	-	-	-	-

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 14. Specificity means by experimental condition and final ANOVA model results for factor score estimate residual methods for Study 1 examining extreme intercept and slope trajectory aberrance.

		Factor Score Type									
		Regression					Bartlett's				
Method	Predictor	M - Low*	M - High*	DF	F	η^2	M - Low*	M - High*	DF	F	η^2
FS Residuals	Sample size	.94	.94	1	39.98	.00	.98	.98	1	24.63	.00
	Obs. times	.94	.94	1	14.00	.00	.99	.98	1	4883.84	.11
	Communality	.94	.93	1	468.67	.00	.98	.98	1	.42	.00
	% Aberrant	.97	.90	1	100676.00	.92	1.00	.97	1	30907.10	.67
	Obs. times*%Aberrant	-	-	-	-	-	-	-	1	2582.23	.06
$RMSR_i$	Sample size	.95	.95	1	1.27	.00	.94	.94	1	.81	.00
	Obs. times	.95	.94	1	641.61	.00	.94	.94	1	.03	.00
	Communality	.95	.95	1	7.11	.00	.94	.94	1	.02	.00
	% Aberrant	.98	.91	1	368519.00	.98	.98	.90	1	816247.00	.99

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 15. Specificity means by experimental condition and final ANOVA model results for log-likelihood methods for Study 1 examining extreme intercept trajectory aberrance.

Method	Predictor	M - Low*	M - High*	DF	F	η^2
$-2PLL_i$	Sample Size	.96	.96	1	1.40	.00
	Obs. times	.96	.96	1	1119.95	.01
	Communality	.96	.96	1	446.55	.00
	% Aberrant	.99	.93	1	157318.00	.94
IND_CHI_i	Sample Size	.95	.95	1	54.56	.00
	Obs. times	.95	.94	1	695.02	.00
	Communality	.95	.95	1	48.39	.00
	% Aberrant	.98	.91	1	308992.00	.97

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 16. Specificity means by experimental condition and final ANOVA model results for log-likelihood methods for Study 1 examining extreme intercept and slope trajectory aberrance.

Method	Predictor	M - Low*	M - High*	DF	F	η^2
$-2PLL_i$	Sample Size	.97	.97	1	2.40	.00
	Obs. times	.97	.97	1	44.76	.00
	Communality	.96	.98	1	10896.40	.08
	% Aberrant	.99	.95	1	114163.00	.83
	Communality* %Aberrant	-	-	1	4074.34	.03
IND_CHI_i	Sample Size	.95	.95	1	137.59	.00
	Obs. times	.95	.95	1	456.63	.00
	Communality	.95	.95	1	214.00	.00
	% Aberrant	.98	.91	1	337936.00	.97

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 17. Mean number of true positives across conditions for Study 2 examining extreme variability aberrance.

		Percent Aberrant Observations			
		2%		10%	
Method	Factor Score Type	N=200, Aber=4	N=1000, Aber=20	N=200, Aber=20	N=1000, Aber=100
Both FS	Regression	.26	1.01	3.06	13.14
	Bartlett's	.50	2.68	4.48	22.00
Either FS	Regression	1.41	6.60	10.42	48.76
	Bartlett's	1.85	9.35	12.56	61.83
FS Residuals	Regression	3.73	18.69	18.70	93.59
	Bartlett's	3.14	15.96	15.14	77.42
Bonf. FS Residuals	Regression	3.59	18.05	14.73	63.28
	Bartlett's	2.62	12.06	8.76	34.16
RMSR_i	Regression	3.72	18.60	18.71	93.72
	Bartlett's	3.58	17.81	18.52	92.74
$-2PLL_i$	-	3.78	18.93	19.07	95.44
IND_CHI_i	-	.39	4.22	5.36	34.35

Note. Sim N = 2000; N= the sample size in each simulated data set; Aber = the total number of aberrant observations in each simulated data set; Both FS = both factor score estimates were identified as aberrant; Either FS = either the intercept or the slope factor score estimate was identified as aberrant

Table 18. Sensitivity across conditions for Study 2 examining extreme variability aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Both FS	Regression	.10	.11	.05	.00	.60
	Bartlett's	.18	.14	.15	.00	1.00
Either FS	Regression	.42	.22	.41	.00	1.00
	Bartlett's	.54	.20	.55	.00	1.00
FS Residuals	Regression	.93	.09	.99	.25	1.00
	Bartlett's	.78	.22	.85	.00	1.00
Bonf. FS Residuals	Regression	.79	.22	.88	.10	1.00
	Bartlett's	.51	.34	.61	.00	1.00
RMSR_i	Regression	.93	.09	.98	.25	1.00
	Bartlett's	.91	.11	.95	.25	1.00
$-2PLL_i$	-	.95	.08	1.00	.25	1.00
IND_CHI_i	-	.23	.14	.25	.00	.75

Notes. Sim N = 8000; Both FS = both factor score estimates were identified as aberrant; Either FS = either the intercept or the slope factor score estimate was identified as aberrant;

Table 19. Specificity across conditions for Study 2 examining extreme variability aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Both FS	Regression	.98	.02	.99	.89	1.00
	Bartlett's	1.00	.00	1.00	.97	1.00
Either FS	Regression	.93	.05	.96	.81	.99
	Bartlett's	.92	.05	.96	.81	1.00
FS Residuals	Regression	1.00	.00	1.00	.96	1.00
	Bartlett's	1.00	.00	1.00	.99	1.00
Bonf. FS Residuals	Regression	1.00	.00	1.00	1.00	1.00
	Bartlett's	1.00	.00	1.00	1.00	1.00
RMSR_i	Regression	1.00	.01	1.00	.95	1.00
	Bartlett's	1.00	.01	1.00	.95	1.00
$-2PLL_i$	-	1.00	.01	1.00	.96	1.00
IND_CHI_i	-	.95	.03	.97	.89	1.00

Notes. Sim N = 8000; Both FS = both factor score estimates were identified as aberrant; Either FS = either the intercept or the slope factor score estimate was identified as aberrant;

Table 20. Sensitivity means by experimental condition and final ANOVA model results for approaches examining factor score estimates directly and factor score estimate residual based approaches for Study 2 examining extreme variability aberrance.

		Factor Score Type									
		Regression					Bartlett's				
Method	Predictor	M - Low*	M - High*	DF	F	η^2	M - Low*	M - High*	DF	F	η^2
Both FS	Sample Size	.11	.09	1	96.96	.01	.18	.18	1	1.08	.00
	Obs. Times	.13	.07	1	1079.62	.07	.21	.14	1	1285.04	.07
	Communality	.16	.05	1	3683.88	.24	.27	.09	1	7261.28	.39
	%Aberrant	.06	.14	1	2197.60	.14	.13	.22	1	1908.53	.10
	Obs. times*Communality	-	-	1	176.66	.01	-	-	-	-	-
	Communality *%Aberrant	-	-	1	215.96	.01	-	-	-	-	-
Either FS	Sample Size	.44	.41	1	52.86	.00	.55	.54	1	.42	.00
	Obs. Times	.49	.36	1	1158.23	.09	.61	.48	1	1955.62	.12
	Communality	.50	.34	1	1756.41	.14	.64	.45	1	3942.46	.24
	%Aberrant	.34	.50	1	1781.58	.14	.46	.62	1	2772.90	.17
	Obs. times*Communality	-	-	1	147.92	.01	-	-	-	-	-
FS Residuals	Sample size	.93	.94	1	1.12	.00	.77	.79	1	46.21	.00
	Obs. Times	.88	.99	1	4995.86	.38	.59	.97	1	28770.80	.78
	Communality	.94	.93	1	4.26	.00	.77	.79	1	48.24	.00
	% Aberrant	.93	.94	1	1.57	.00	.79	.77	1	130.37	.00
$RMSR_i$	Sample size	.93	.93	1	.34	.00	.91	.91	1	.87	.00
	Obs. Times	.88	.99	1	4106.95	.34	.84	.98	1	6581.87	.43
	Communality	.94	.93	1	43.09	.00	.91	.91	1	2.75	.00
	% Aberrant	.93	.94	1	14.88	.00	.89	.93	1	381.34	.02
	Obs. times*%Aberrant	-	-	-	-	-	-	-	1	324.57	.02

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 21. Sensitivity means by experimental condition and final ANOVA model results for log-likelihood methods for Study 2 examining extreme variability aberrance.

Method	Predictor	M - Low*	M - High*	DF	F	η^2
$-2PLL_i$	Sample size	.95	.95	1	.67	.00
	Obs. times	.91	.99	1	3397.78	.28
	Communality	.97	.94	1	429.38	.04
	% Aberrant	.95	.95	1	30.11	.00
	Obs. times*Communality	-	-	1	322.37	.03
IND_CHI_i	Sample size	.18	.28	1	1810.86	.11
	Obs. times	.27	.19	1	1432.15	.09
	Communality	.23	.23	1	.34	.00
	% Aberrant	.15	.31	1	4653.27	.28
	Obs. times*Sample size	-	-	1	307.52	.02
	Obs. times*%Aberrant	-	-	1	239.76	.01

* Low group is sample size of 200, 4 time points, .4 communality, 2% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 10% aberrant observations

Note. These ANOVA models were fit to 8000 simulated data sets

Table 22. Mean number of true positives across conditions for Study 3 examining functional form aberrance.

		Sample Size					
		200			1000		
Method	Factor Score Type	10%, Aber= 20	25%, Aber= 50	40%, Aber= 80	10%, Aber= 100	25%, Aber= 250	40%, Aber= 400
Quad FS	Regression	2.26	4.91	6.34	9.55	21.85	27.57
	Bartlett's	4.08	8.34	10.90	19.18	38.92	50.72
FS Residuals	Regression	7.26	18.29	30.40	39.22	100.44	168.39
	Bartlett's	5.99	13.18	19.32	29.96	65.67	97.08
Bonf. FS Residuals	Regression	4.17	9.28	14.71	20.30	44.10	68.13
	Bartlett's	3.14	5.71	7.29	13.76	22.79	28.28

Note. Sim N =4000 number of simulation observations used to calculate the mean number of true positives; N= the sample size in each simulated data set; Aber = the total number of aberrant observations in each simulated data set;

Table 23. Sensitivity across conditions for Study 3 examining functional form aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Quad FS	Regression	.09	.18	.00	.00	.90
	Bartlett's	.16	.16	.10	.00	.90
FS Residuals	Regression	.39	.24	.39	.00	1.00
	Bartlett's	.27	.17	.20	.00	1.00
Bonf. FS Residuals	Regression	.19	.19	.12	.00	1.00
	Bartlett's	.11	.13	.05	.00	.85

Note. Sim N =24000

Table 24. Specificity across conditions for Study 3 examining functional form aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Quad FS	Regression	.99	.03	1.00	.50	1.00
	Bartlett's	.94	.02	.94	.77	.99
FS Residuals	Regression	.86	.12	.89	.17	1.00
	Bartlett's	.90	.04	.88	.76	1.00
Bonf. FS Residuals	Regression	.97	.08	.99	.21	1.00
	Bartlett's	.98	.01	.99	.90	1.00

Note. Sim N =24000

Table 25. AUC across conditions for Study 3 examining functional form aberrance.

Method	Factor Score Type	M	SD	Median	Min	Max
Quad FS	Regression	.69	.15	.64	.44	1.00
	Bartlett's	.68	.14	.65	.47	1.00
FS Residuals	Regression	.62	.13	.57	.39	1.00
	Bartlett's	.67	.14	.63	.39	1.00
RMSR_i	Regression	.62	.13	.57	.39	1.00
	Bartlett's	.66	.14	.61	.41	1.00
$-2PLL_i$	-	.59	.10	.56	.40	.99

Note. Sim N =24000

Table 26. AUC means by experimental condition and final ANOVA model results for methods examining factor score estimates for Study 3 examining functional form aberrance.

		Factor Score Type											
		Regression						Bartlett's					
Method	Predictor	M - Low*	M - Med*	M - High*	DF	F	η^2	M - Low*	M - Med*	M - High*	DF	F	η^2
Quad FS	Sample size	.69	-	.69	1	38.52	.00	.68	-	.68	1	101.5	.00
	Obs. Times	.59	-	.79	1	112010	.49	.58	-	.78	1	172290	.52
	Communality	.65	-	.73	1	16730.6	.07	.63	-	.73	1	39440.7	.12
	% Aberrant	.69	.69	.69	2	22.1	.00	.68	.68	.68	2	52.3	.00
	Quadratic Size	.62	-	.76	1	56781.3	.24	.62	-	.75	1	73259.2	.22
	Obs. times* Communality	-	-	-	-	-	-	-	-	-	1	5387.9	.02
	Obs. times* Quadratic size	-	-	-	1	20243.8	.09	-	-	-	1	16243.4	.05

* Low group is sample size of 200, 4 time points, .4 communality, 10% aberrant observations; Medium group has 25% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 40% aberrant observations

Note. These ANOVA models were fit to 24000 simulated data sets

Table 27. AUC means by experimental condition and final ANOVA model results for methods examining factor score residual approaches for Study 3 examining functional form aberrance.

		Factor Score Type											
		Regression						Bartlett's					
Method	Predictor	M - Low*	M - Med*	M - High*	DF	F	η^2	M - Low*	M - Med*	M - High*	DF	F	η^2
FS Residuals	Sample size	.67	-	.68	1	228.5	.00	.63	-	.62	1	248.3	.00
	Obs. Times	.57	-	.77	1	115051	.48	.53	-	.71	1	125982	.45
	Communality	.63	-	.71	1	19495.3	.08	.58	-	.67	1	33369.7	.12
	% Aberrant	.67	.67	.68	2	75.9	.00	.63	.62	.62	2	182.3	.00
	Quadratic Size	.60	-	.74	1	59333.8	.25	.57	-	.68	1	47242.3	.17
	Obs. times* Communality	-	-	-	1	3661.2	.02	-	-	-	1	20327.6	.07
	Obs. times* Quadratic size	-	-	-	1	19441.0	.08	-	-	-	1	30223.5	.11
$RMSR_i$	Sample size	.66	-	.65	1	9.39	.00	.62	-	.61	1	429.8	.00
	Obs. Times	.55	-	.76	1	161140	.53	.54	-	.70	1	133980	.43
	Communality	.61	-	.70	1	27361.1	.09	.58	-	.66	1	37689.4	.12
	% Aberrant	.66	.65	.66	2	3.2	.00	.63	.62	.61	2	184.0	.00
	Quadratic Size	.60	-	.72	1	54477.5	.18	.57	-	.67	1	53924.1	.17
	Obs. times* Communality	-	-	-	1	8932.0	.03	-	-	-	1	24314.7	.08
	Obs. times* Quadratic size	-	-	-	1	25994.0	.09	-	-	-	1	36970.6	.12
	Communality* Quadratic size	-	-	-	-	-	-	-	-	-	1	3753.0	.01

* Low group is sample size of 200, 4 time points, .4 communality, 10% aberrant observations; Medium group has 25% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 40% aberrant observations

Note. These ANOVA models were fit to 24000 simulated data sets

Table 28. AUC means by experimental condition and final ANOVA model results for the -2PLL_i approach in Study 3 examining functional form aberrance.

Predictor	M - Low*	M – Med*	M - High*	DF	F	η^2
Sample size	.60	-	.59	1	222.35	.00
Obs. times	.54	-	.65	1	59201.60	.34
Communality	.57	-	.62	1	12072.90	.07
% Aberrant	.63	.60	.56	2	7322.71	.08
Quadratic size	.56	-	.63	1	24090.10	.14
Obs. times* Communality	-	-	-	1	5868.73	.03
Obs. times* % Aberrant	-	-	-	2	3845.75	.04
Obs. times* Quadratic size	-	-	-	1	16076.30	.09
% Aberrant* Quadratic size	-	-	-	2	2708.82	.03
Obs. times*% Aberrant* Quadratic size	-	-	-	2	2107.94	.02

* Low group is sample size of 200, 4 time points, .4 communality, 10% aberrant observations; Medium group has 25% aberrant observations; High group is sample size of 1000, 8 time points, .8 communality, 40% aberrant observations

Note. These ANOVA models were fit to 24000 simulated data sets

Figure 1. Mean sensitivity for detecting extreme intercept aberrance using the regression factor score estimate approach as a function of number of observation times and communality.

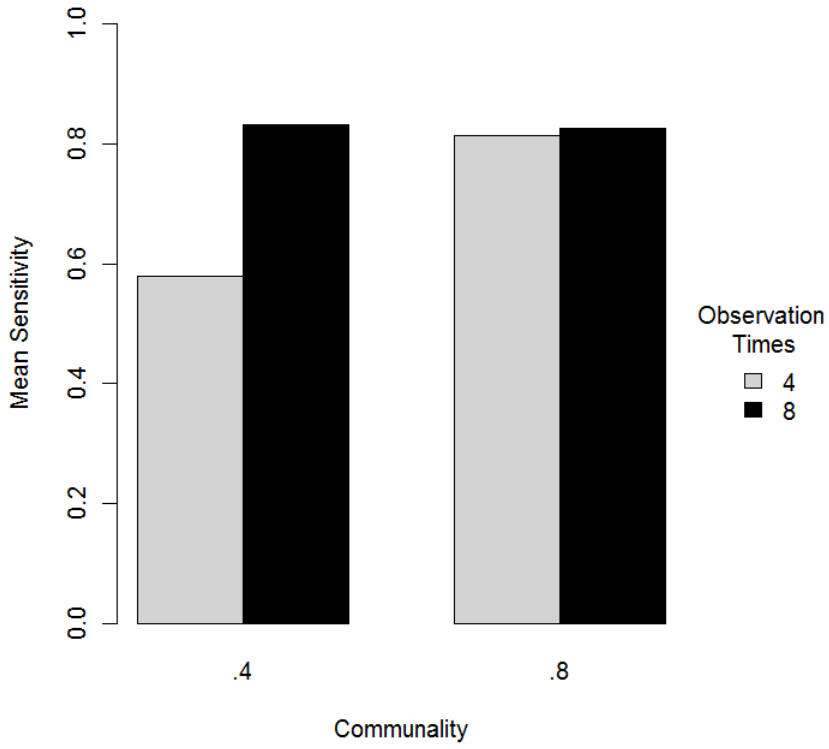


Figure 2. Mean sensitivity for detecting extreme intercept aberrance using the Bartlett's factor score estimate approach as a function of number of observation times and communality.

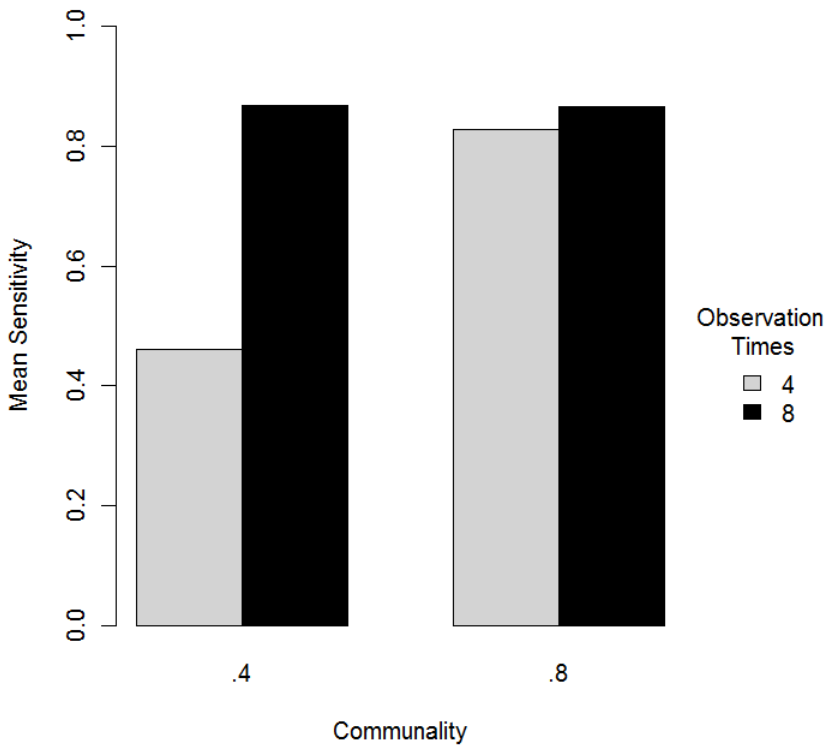


Figure 3. Mean sensitivity for detecting extreme intercept aberrance using Bartlett's factor score estimate approach as a function of percent of aberrant observations and number of observation times.

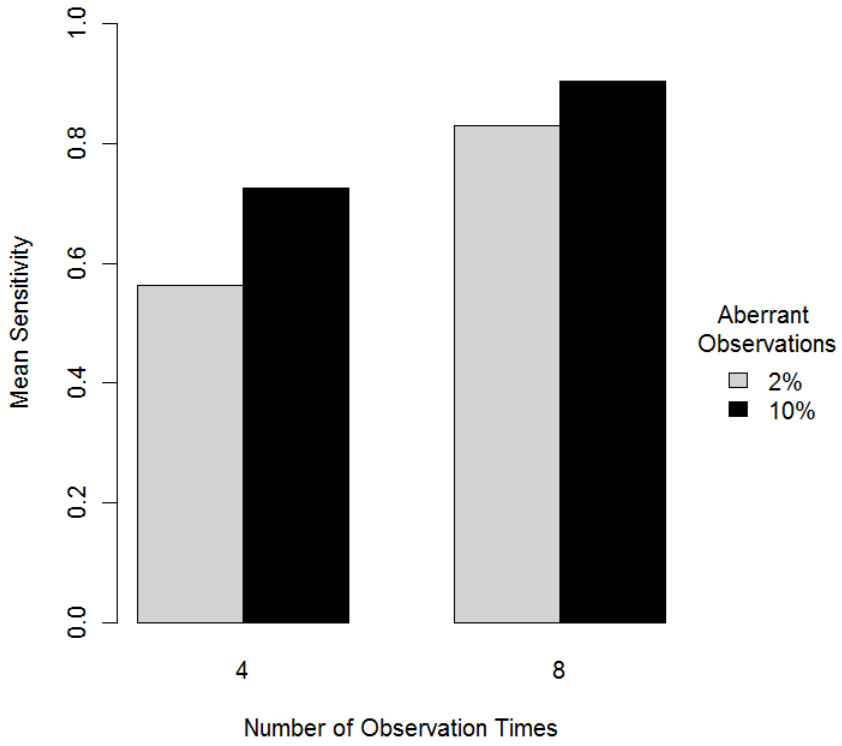


Figure 4. Mean sensitivity for detecting extreme intercept and slope aberrance using the approach examining both regression factor score estimates as a function of communality, number of observation times and percent of aberrant observations.

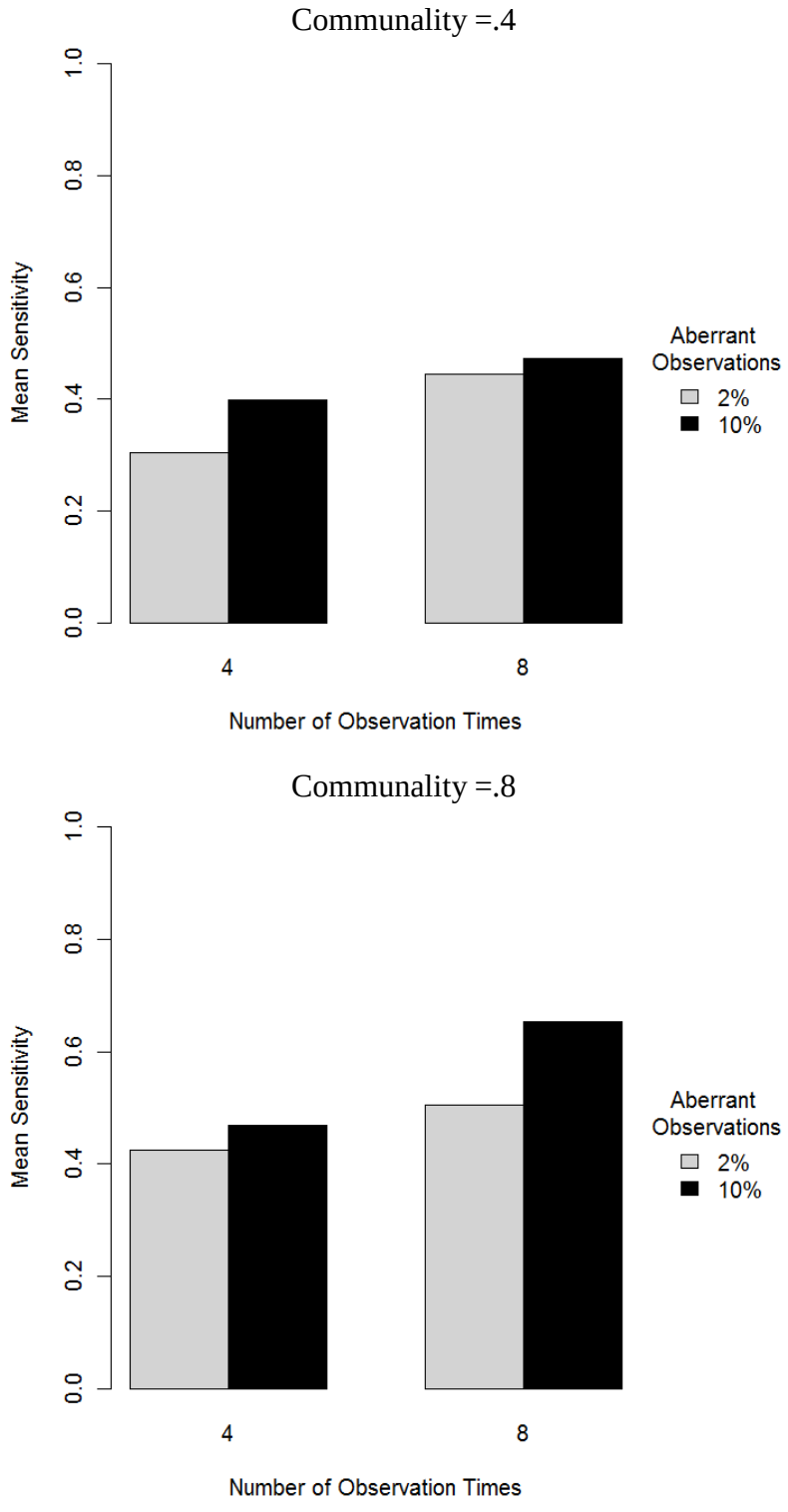


Figure 5. Mean sensitivity for detecting extreme intercept and slope aberrance using the approach examining both Bartlett's factor score estimates as a function of number of observation times and communality.

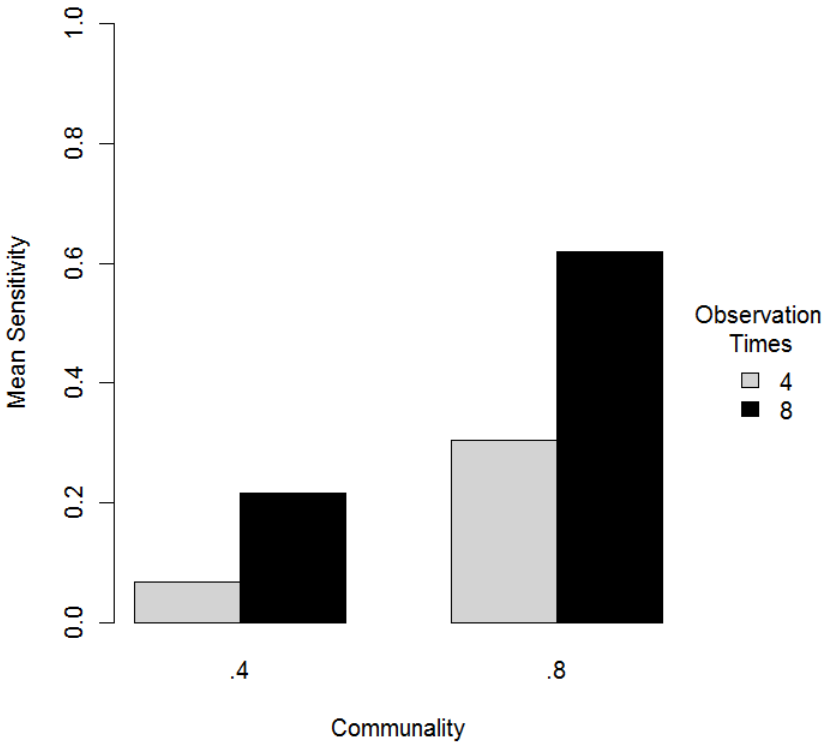


Figure 6. Mean sensitivity for detecting extreme intercept and slope aberrance using the approach examining either regression factor score estimate as a function of number of observation times and communality.

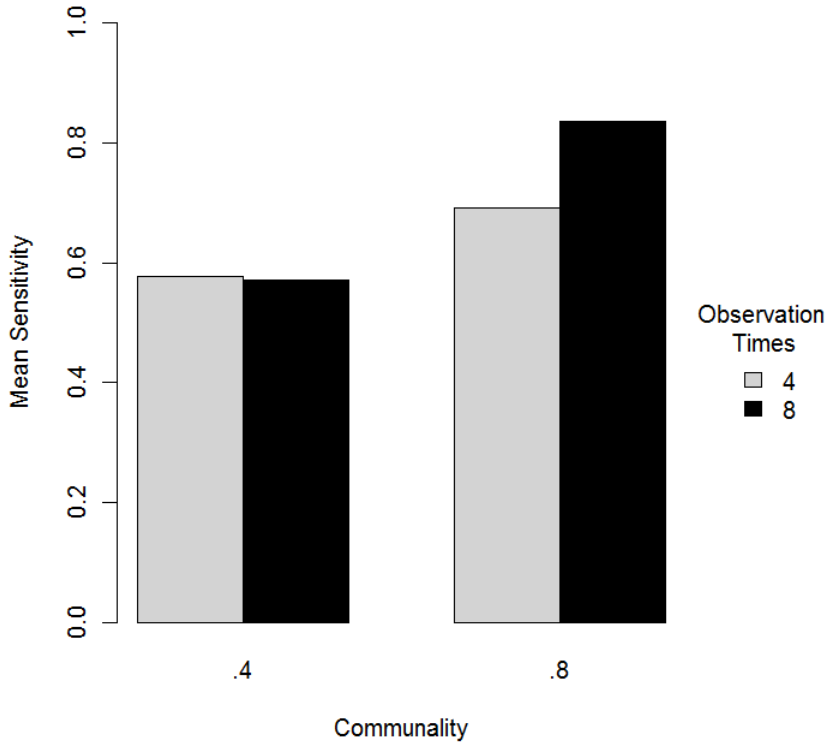


Figure 7. Mean sensitivity for detecting extreme intercept aberrance using the Bartlett's factor score estimate residual approach as a function of percent of aberrant observations and number of observation times.

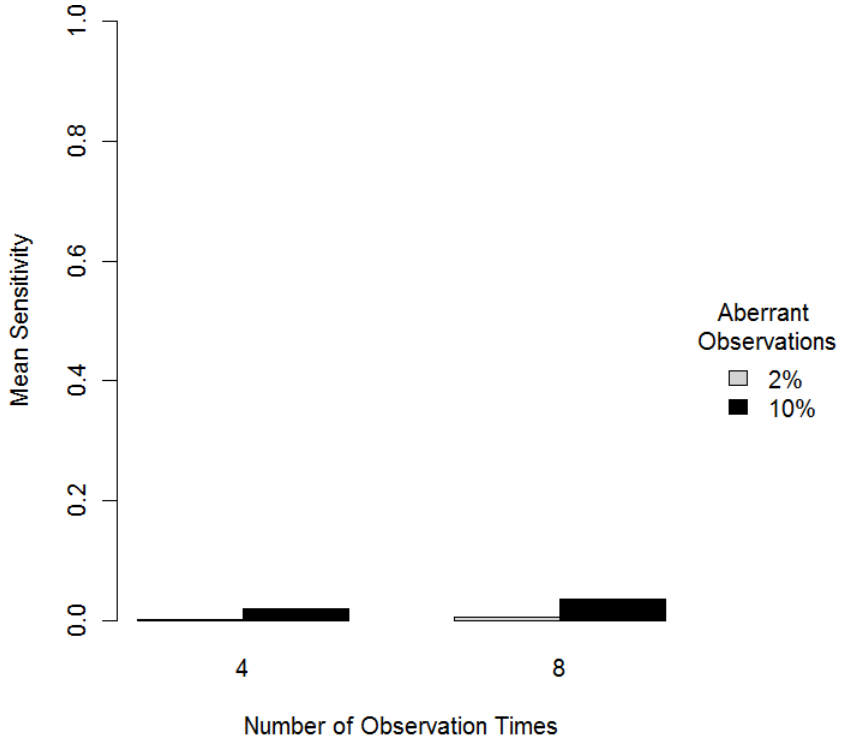


Figure 8. Mean sensitivity for detecting extreme intercept and slope aberrance using the Bartlett's factor score estimate residual approach as a function of percent of aberrant observations and number of observation times.

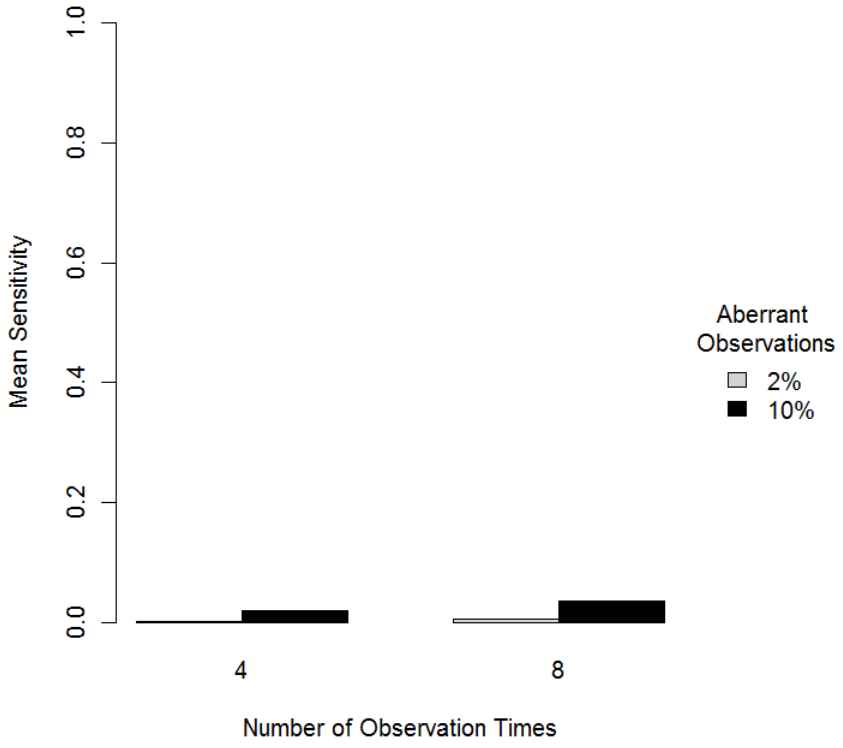


Figure 9. Mean sensitivity for detecting extreme intercept and slope aberrance using the regression factor score residual analysis approach as a function of number of observation times and communality.

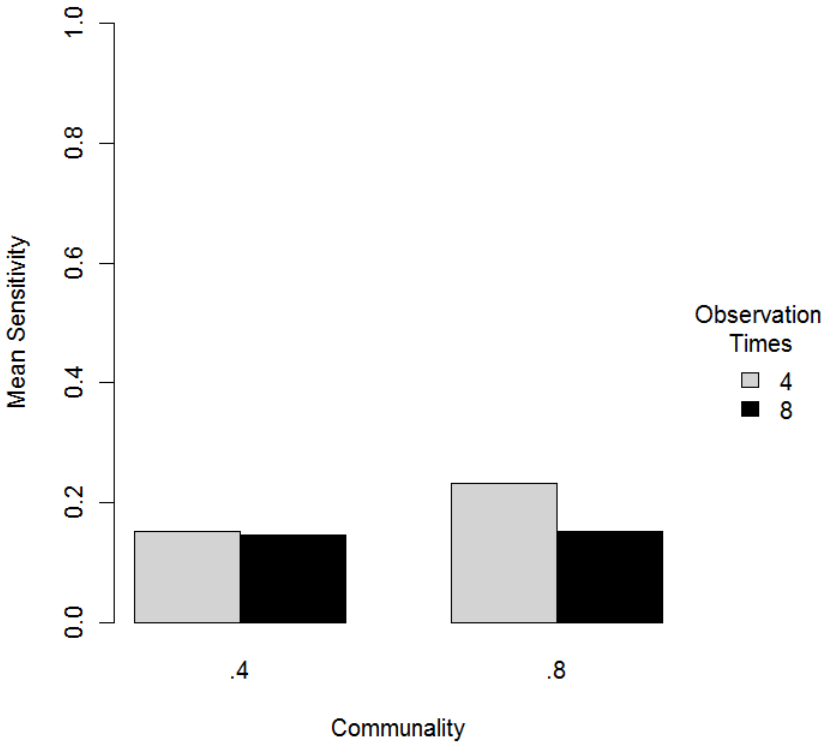


Figure 10. Mean sensitivity for detecting extreme intercept and slope aberrance using the regression $RMSR_i$ approach as a function of number of observation times and communality.

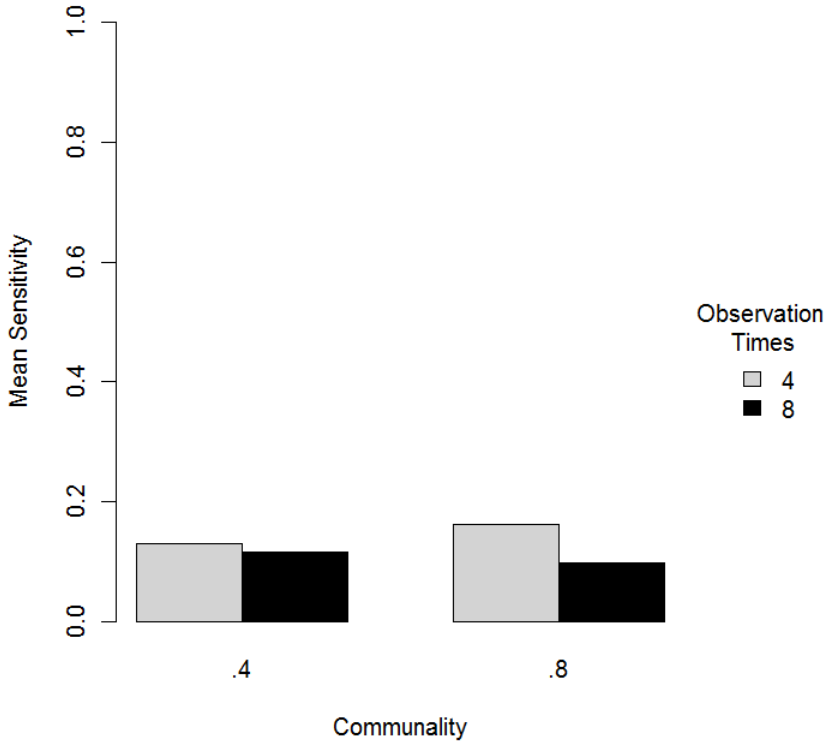


Figure 11. Mean sensitivity for detecting extreme intercept aberrance using the -2PLL_i approach as a function of number of observation times and communality.

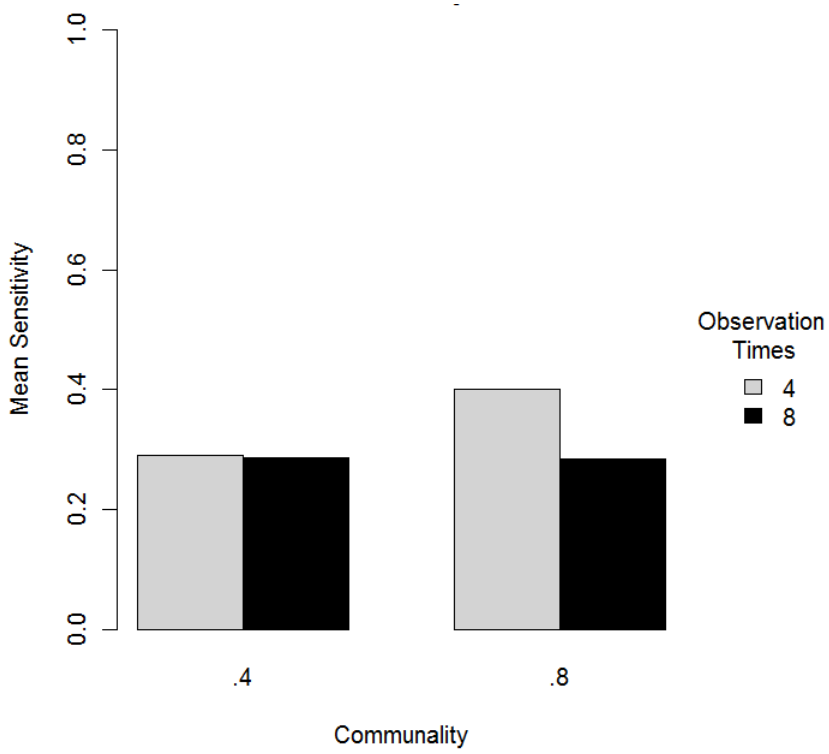


Figure 12. Mean sensitivity for detecting extreme intercept and slope aberrance using the approach examining either Bartlett's factor score estimates as a function of percent of aberrant observations and communality.

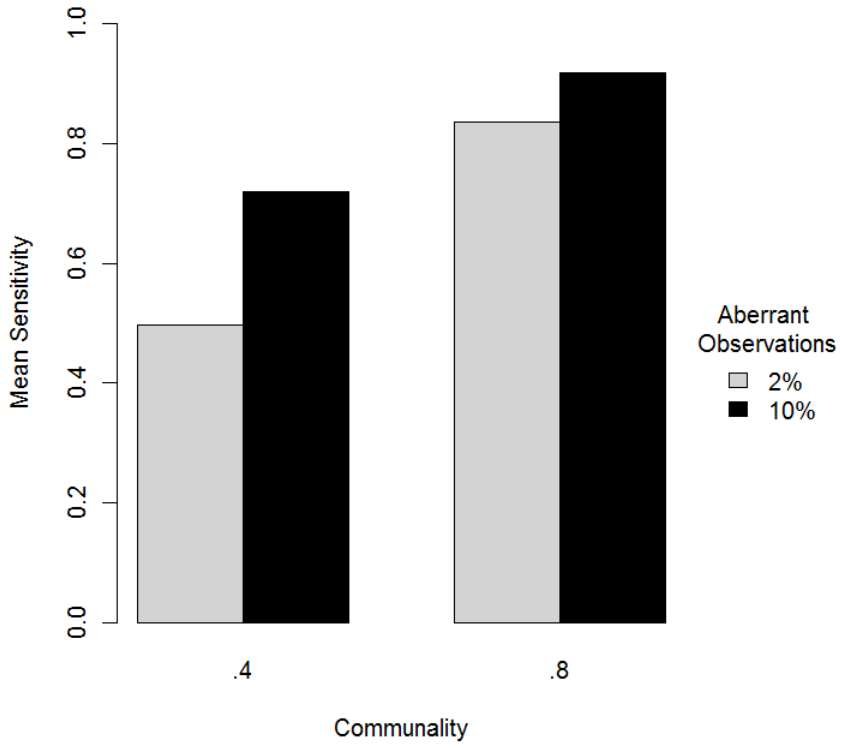
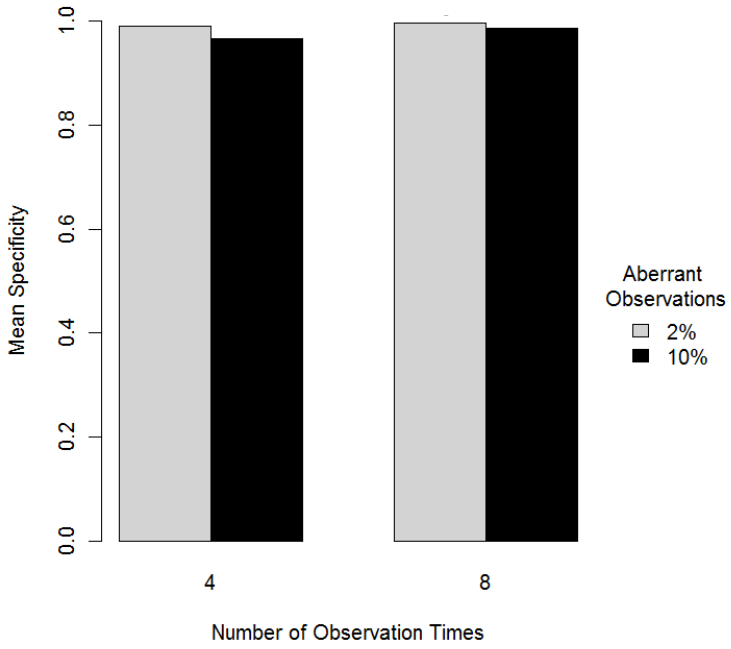


Figure 13. Mean specificity for detecting extreme intercept aberrance using the regression factor score estimate approach as a function of communality, number of observation times and percent of aberrant observations.

Communality =.4



Communality =.8

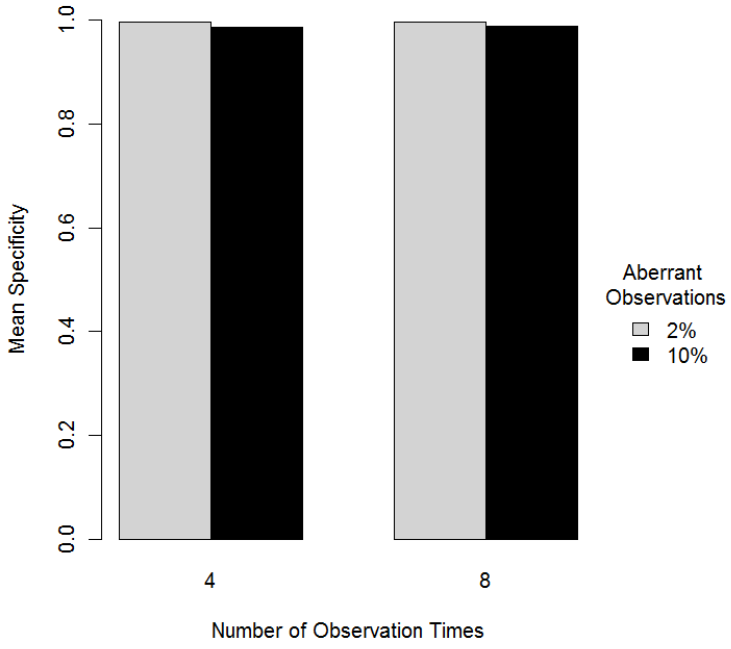
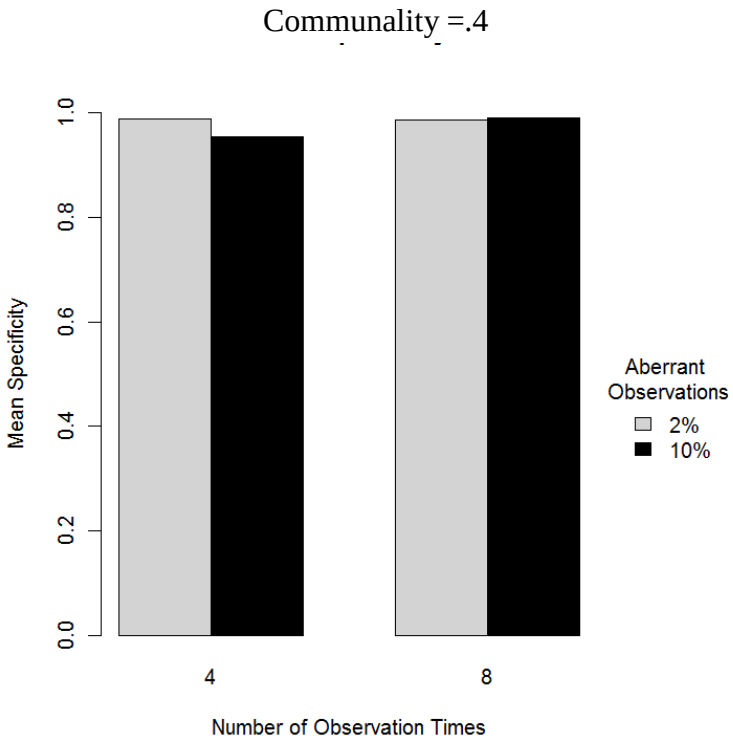


Figure 14. Mean specificity for detecting extreme intercept aberrance using the Bartlett's factor score estimate approach as a function of communality, number of observation times and percent of aberrant observations.



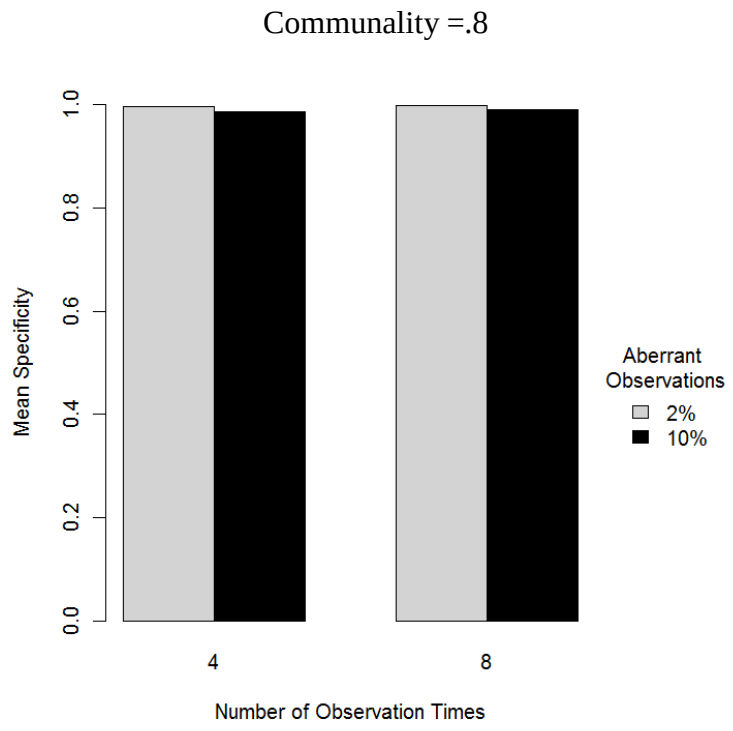
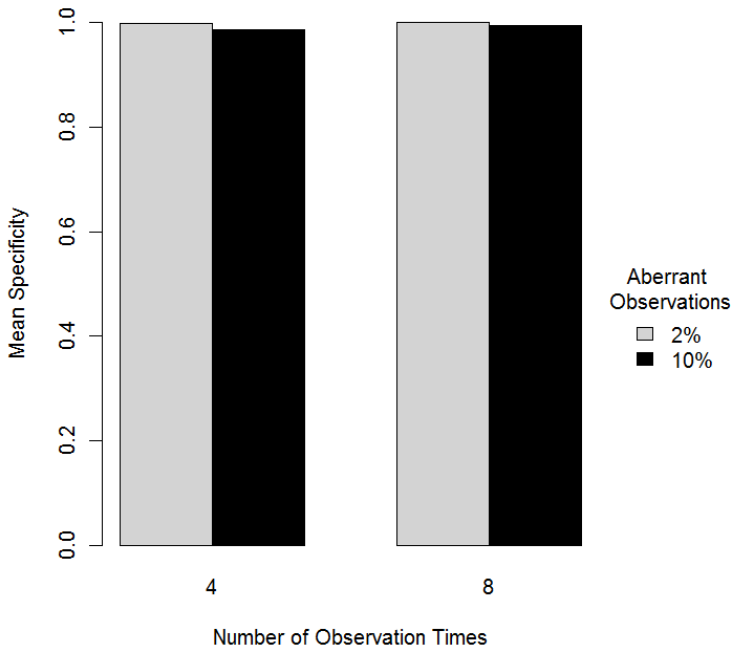


Figure 15. Mean specificity for detecting extreme intercept and slope aberrance using the approach examining both Bartlett's factor score estimates as a function of communality, number of observation times and percent of aberrant observations.

Communality =.4



Communality =.8

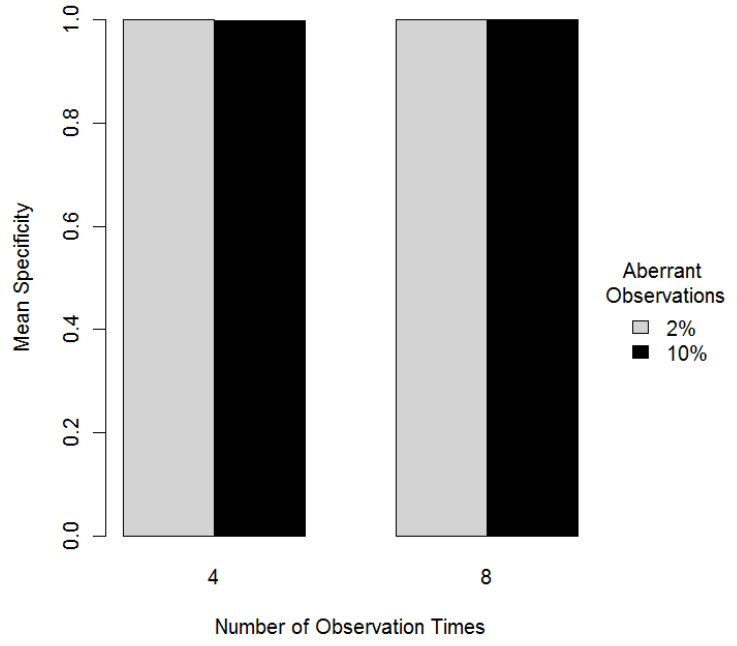
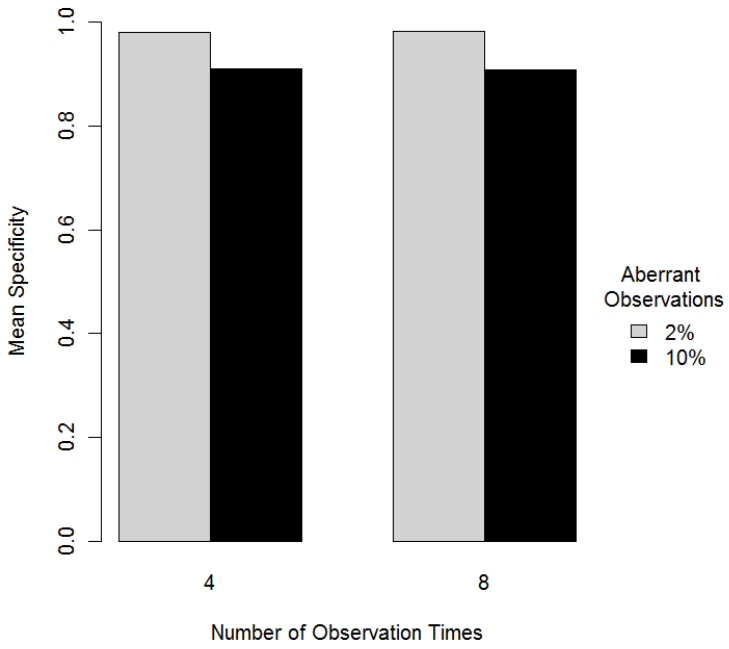


Figure 16. Mean specificity for detecting extreme intercept and slope aberrance using the approach examining either regression factor score estimates as a function of communality, number of observation times and percent of aberrant observations.

Communality =.4



Communality =.8

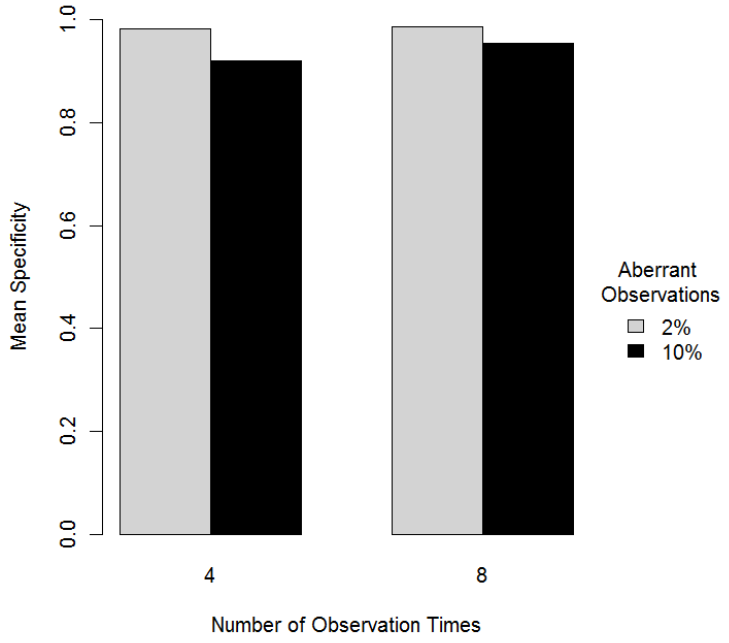


Figure 17. Mean specificity for detecting extreme intercept and slope aberrance using the approach examining both regression factor score estimates as a function of number of observation times and communality.

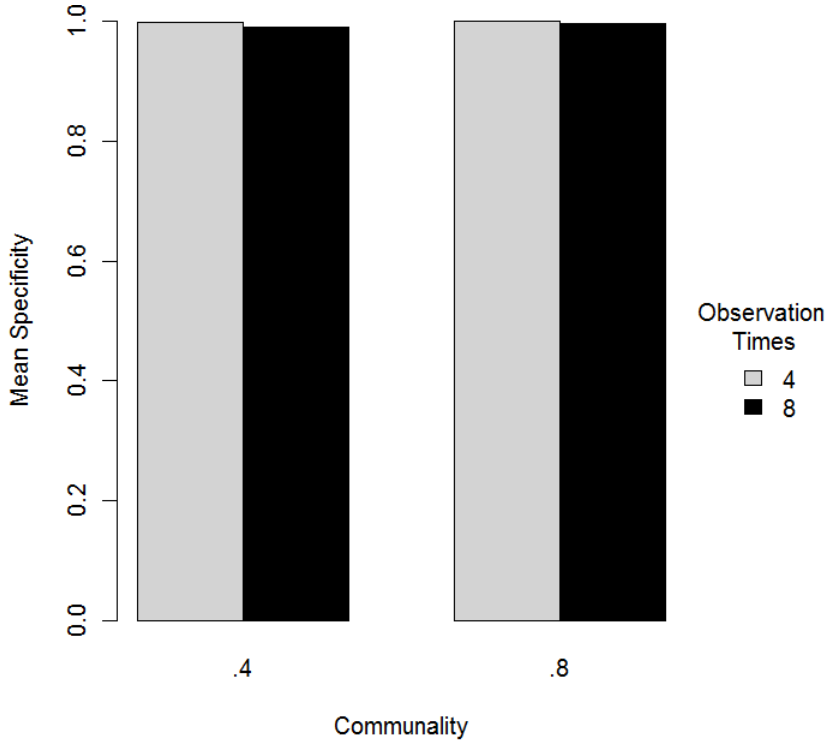


Figure 18. Mean specificity for detecting extreme intercept and slope aberrance using the approach examining both regression factor score estimates as a function of percent of aberrant observations and communality.

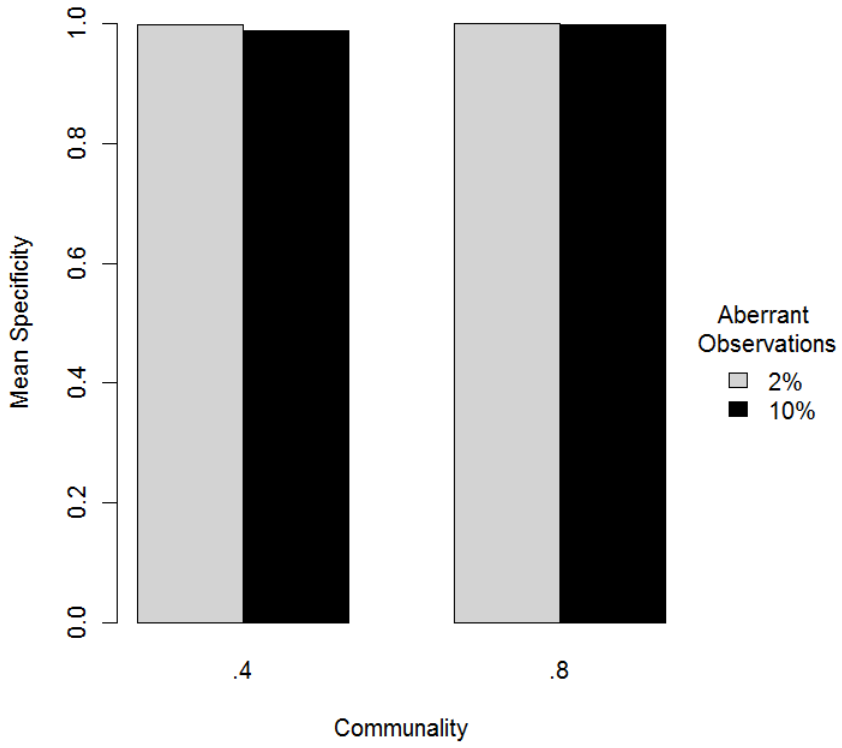


Figure 19. Mean specificity for detecting extreme intercept and slope aberrance using the approach examining either Bartlett's factor score estimates as a function of percent of aberrant observations and number of observation times.

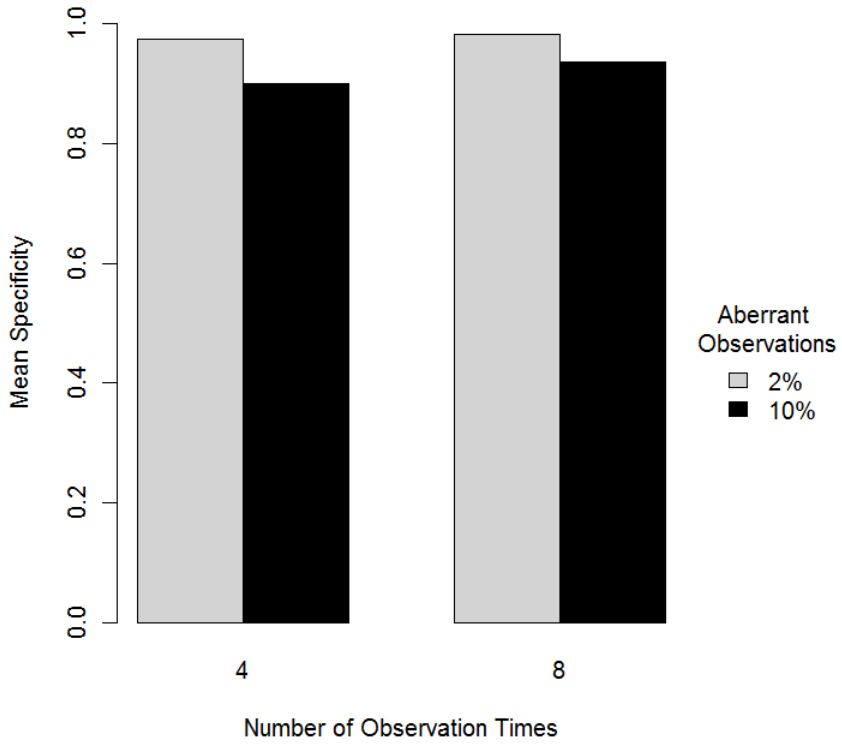


Figure 20. Mean specificity for detecting extreme intercept and slope aberrance using the approach examining either Bartlett's factor score estimates as a function of percent of aberrant observations and communality.

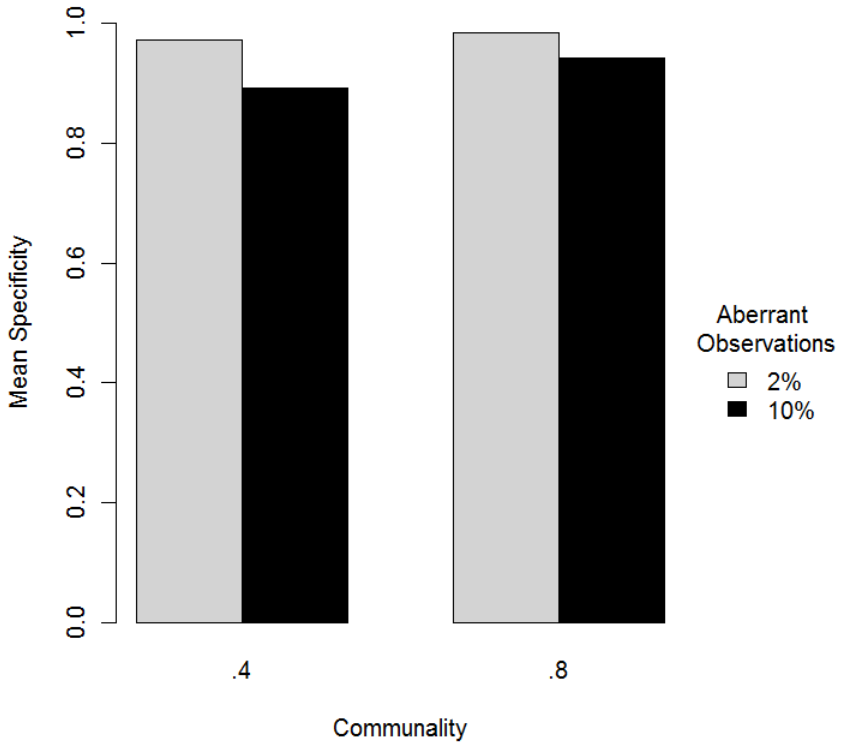


Figure 21. Mean specificity for detecting extreme intercepts using the Bartlett's factor score estimate residual approach as a function of percent of aberrant observations and number of observation times.

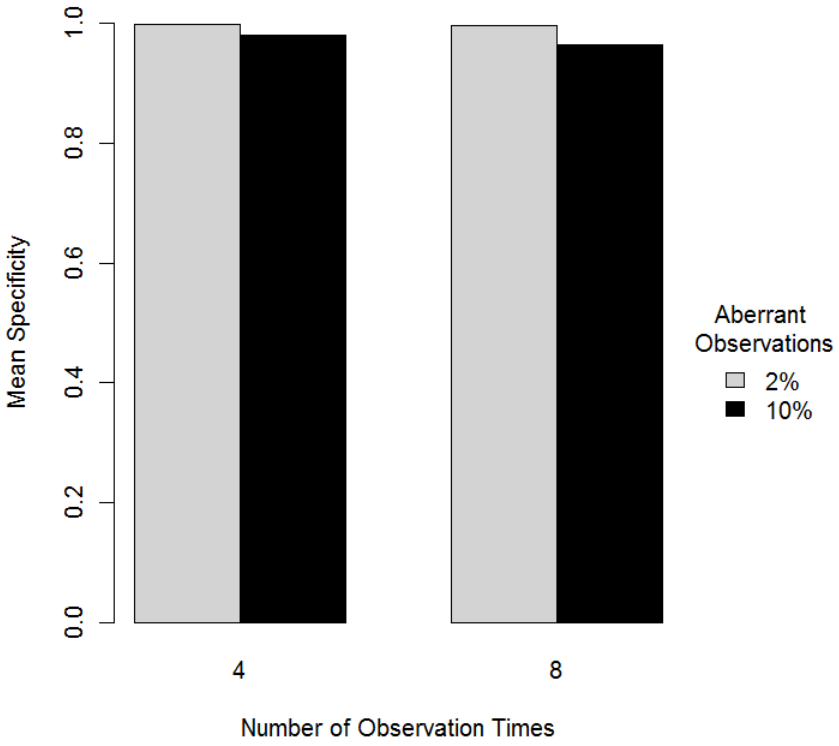


Figure 22. Mean specificity for detecting extreme intercept and slope aberrance using the Bartlett's factor score estimate residual approach as a function of percent of aberrant observations and number of observation times.

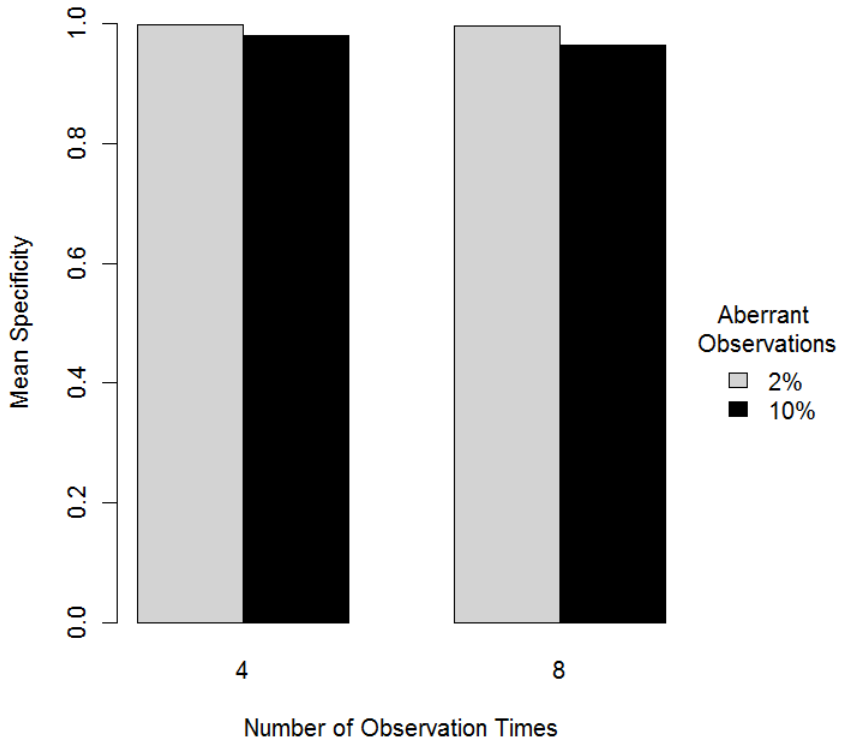


Figure 23. Mean specificity for detecting extreme intercept and slope aberrance using the -2PLL_i approach as a function of percent of aberrant observations and communality.

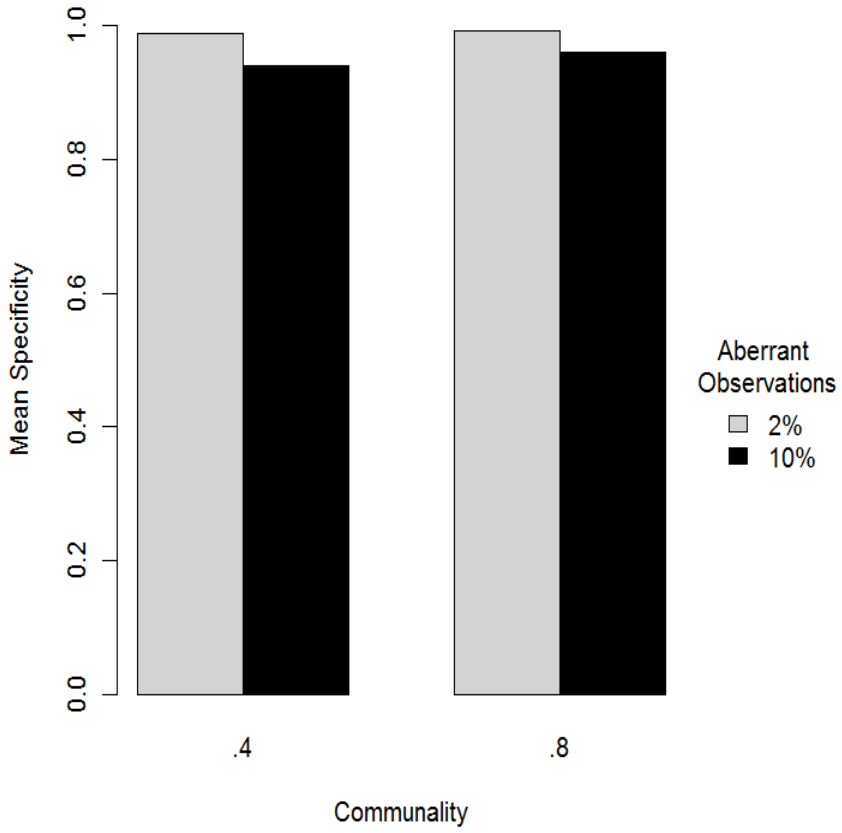


Figure 24. Mean sensitivity for detecting extreme variability aberrance using the approach examining both regression factor score estimates as a function of number of observation times and communality.

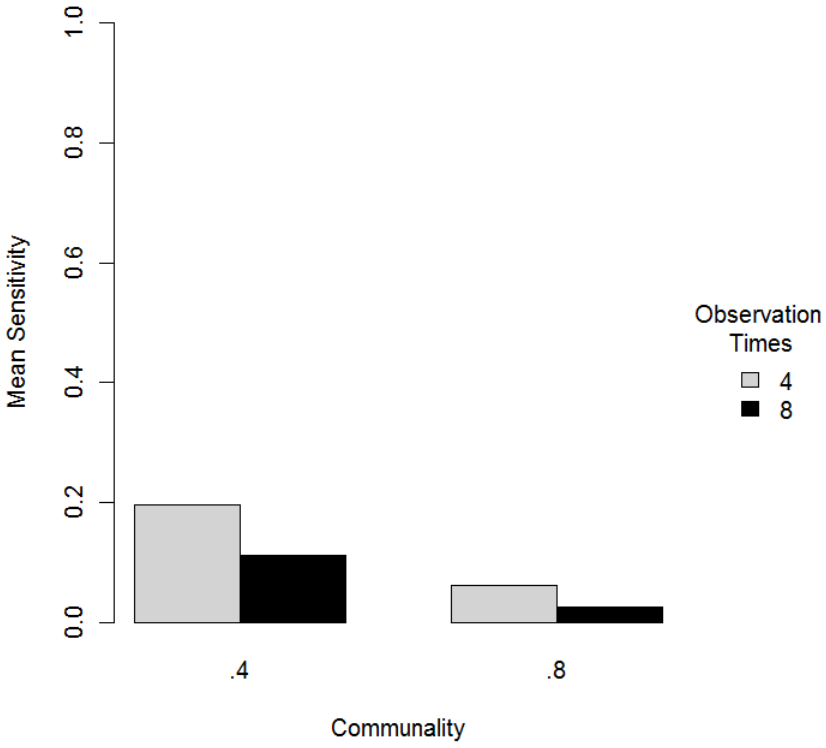


Figure 25. Mean sensitivity for detecting extreme variability aberrance using the approach examining both regression factor score estimates as a function of communality and percent of aberrant observations.

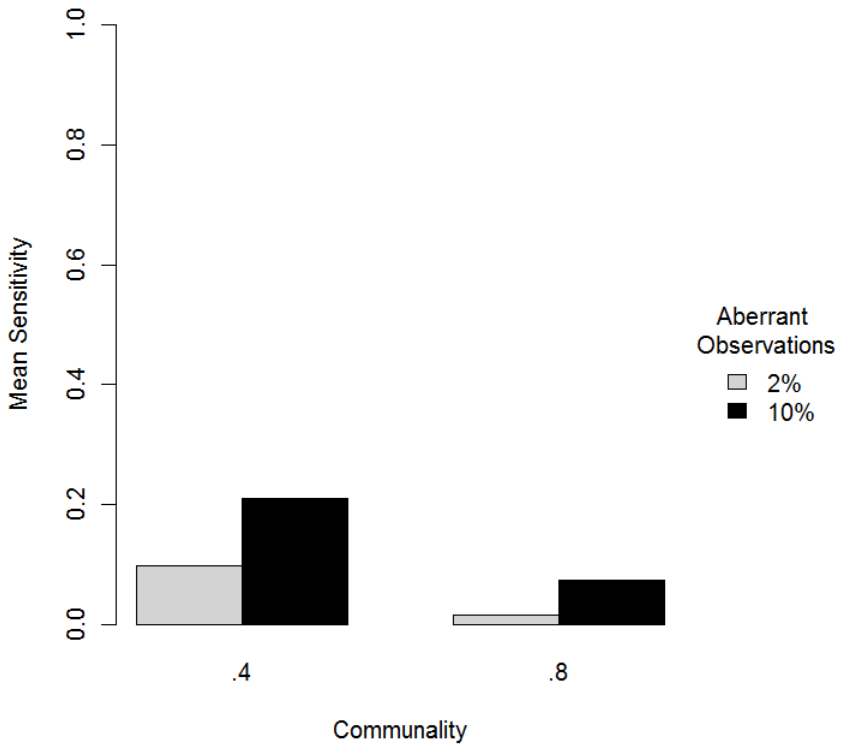


Figure 26. Mean sensitivity for detecting extreme variability aberrance using the approach examining either regression factor score estimate as a function of number of observation times and communality.

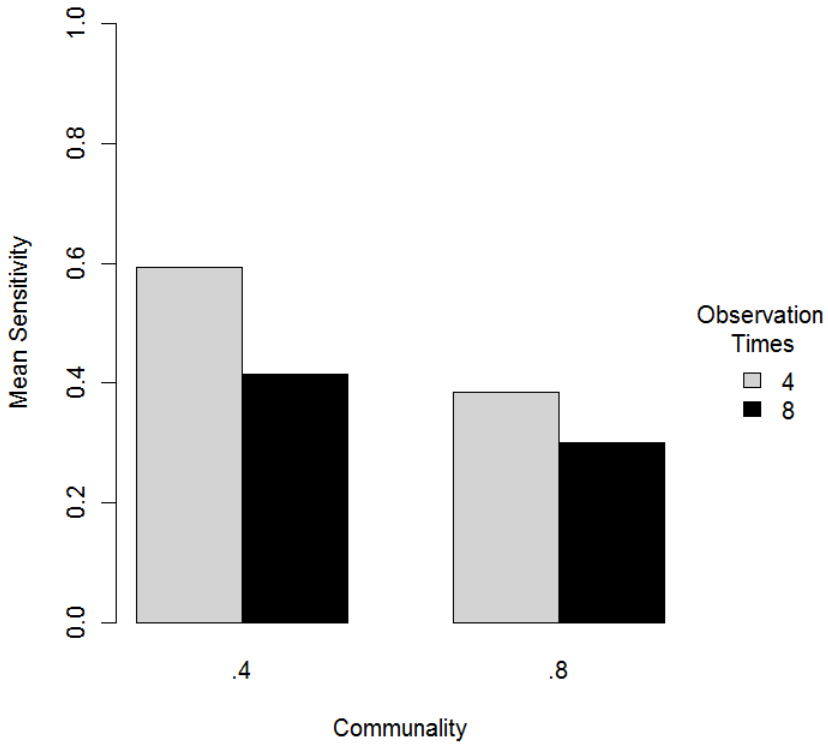


Figure 27. Mean sensitivity for detecting extreme variability aberrance using the Bartlett's $RMSR_i$ approach as a function of number of observation times and percent of aberrant observations.

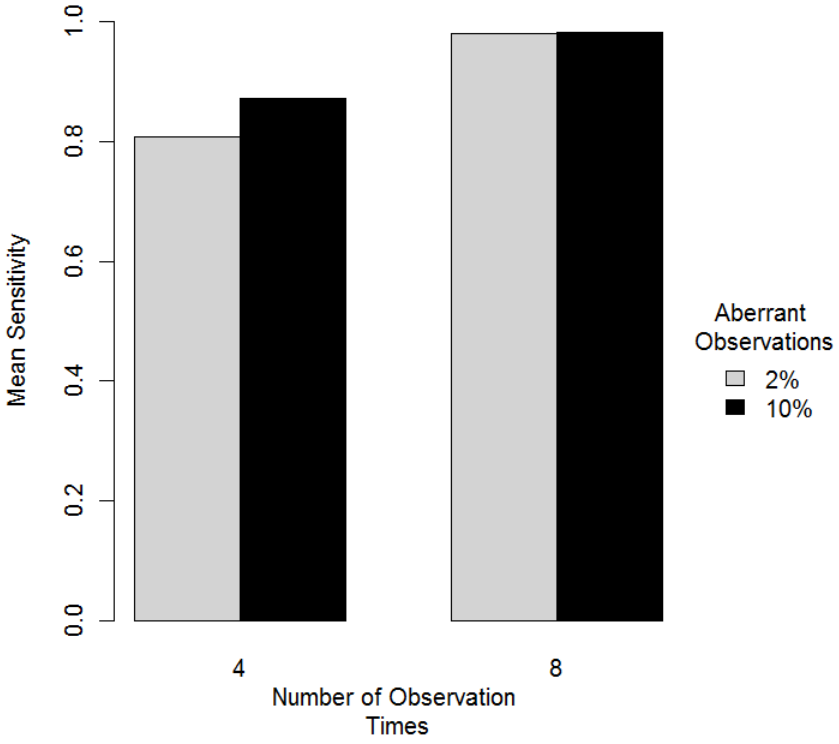


Figure 28. Mean sensitivity for detecting extreme variability aberrance using the -$2PLL_i$ approach as a function of number of observation times and communality.

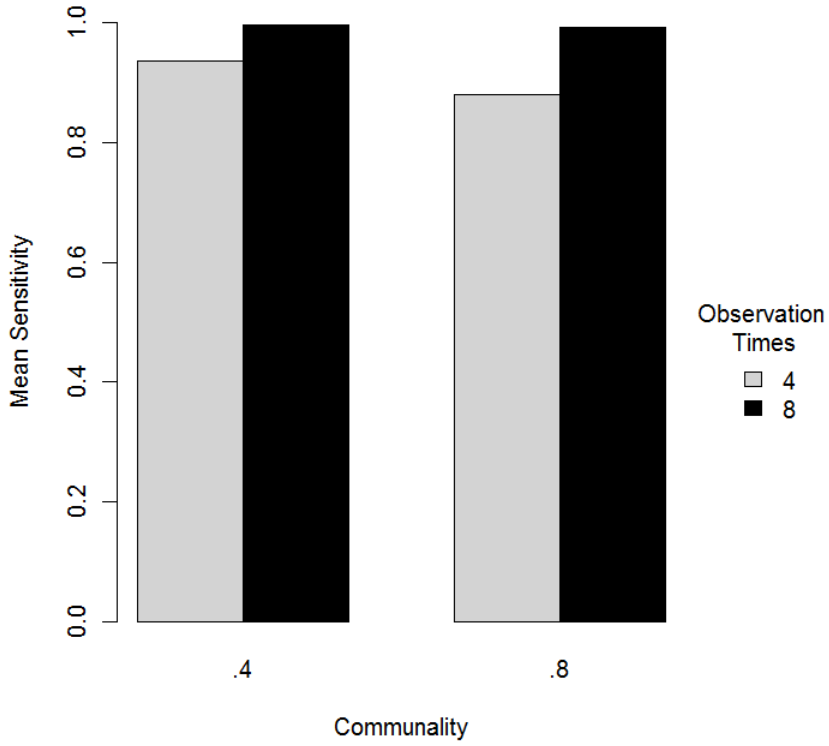


Figure 29. Mean sensitivity for detecting extreme variability aberrance using the IND_CHI_i approach as a function of sample size and number of observation times.

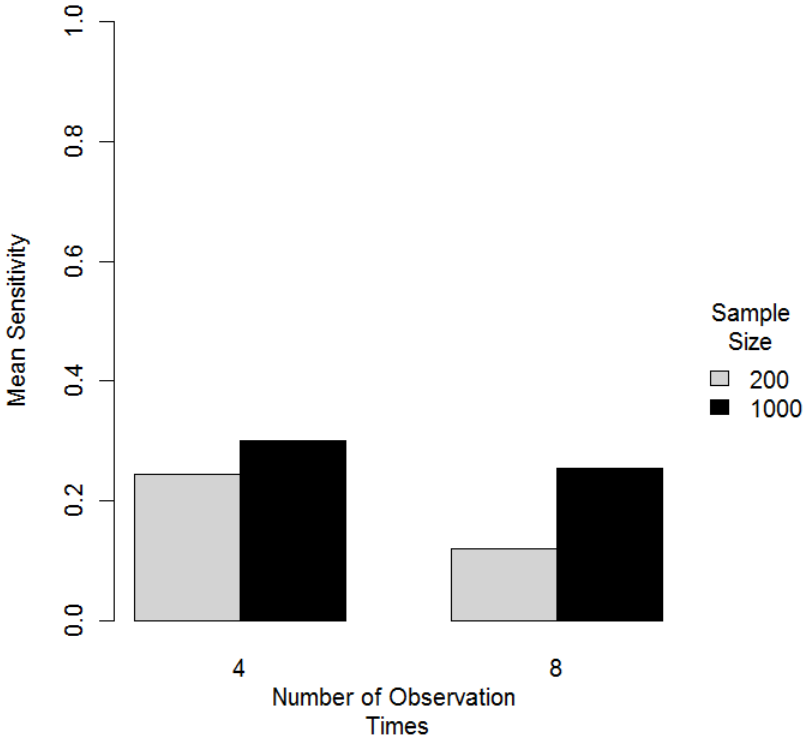


Figure 30. Mean sensitivity for detecting extreme variability aberrance using the IND_CHI_i approach as a function of number of observation times and percent of aberrant observations.

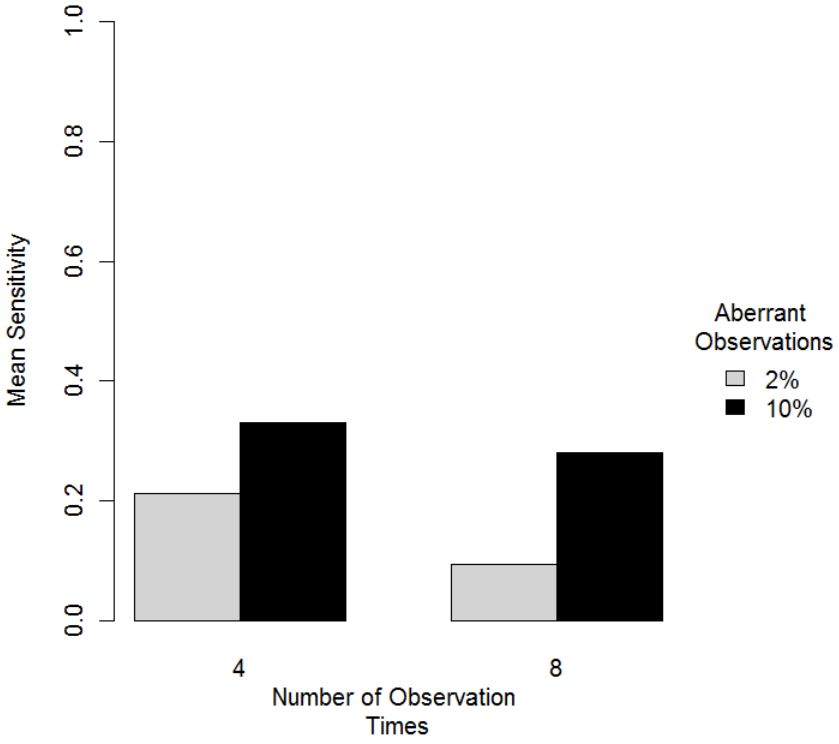


Figure 31. Mean trajectories for reading scores over time for the linear (black), small quadratic (blue), and large quadratic (red) over 8 time points. Note that the large quadratic trajectory may not be a plausible trajectory for reading over time. However, the goal of the simulation is to determine the impact of the size of the difference in trajectories on the selection of an appropriate functional form for a given individual.

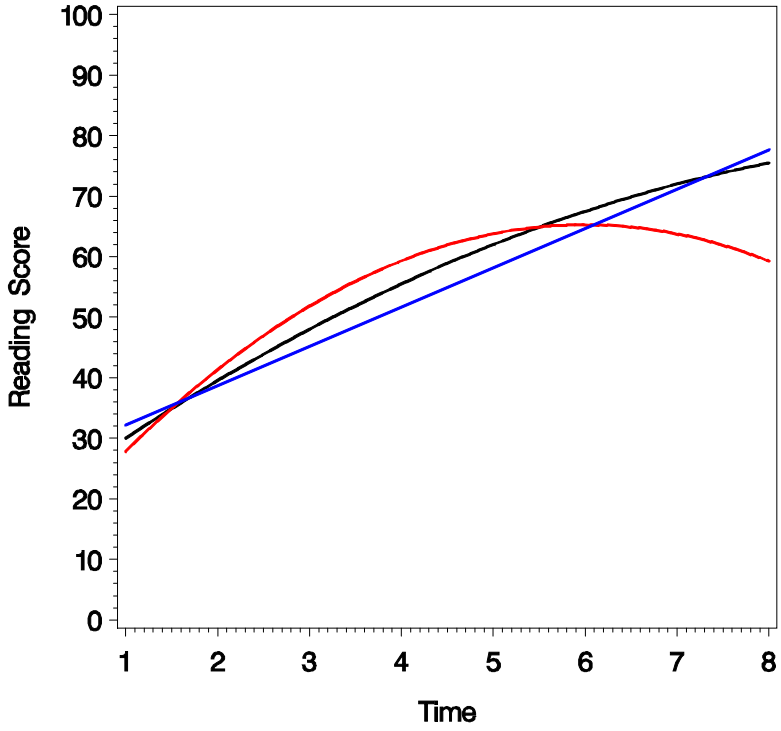


Figure 32. Example ROC plot using the difference in regression factor score residuals approach with an AUC of .40. The grey line represents chance classification and the black line represents the results from a replication with an AUC.

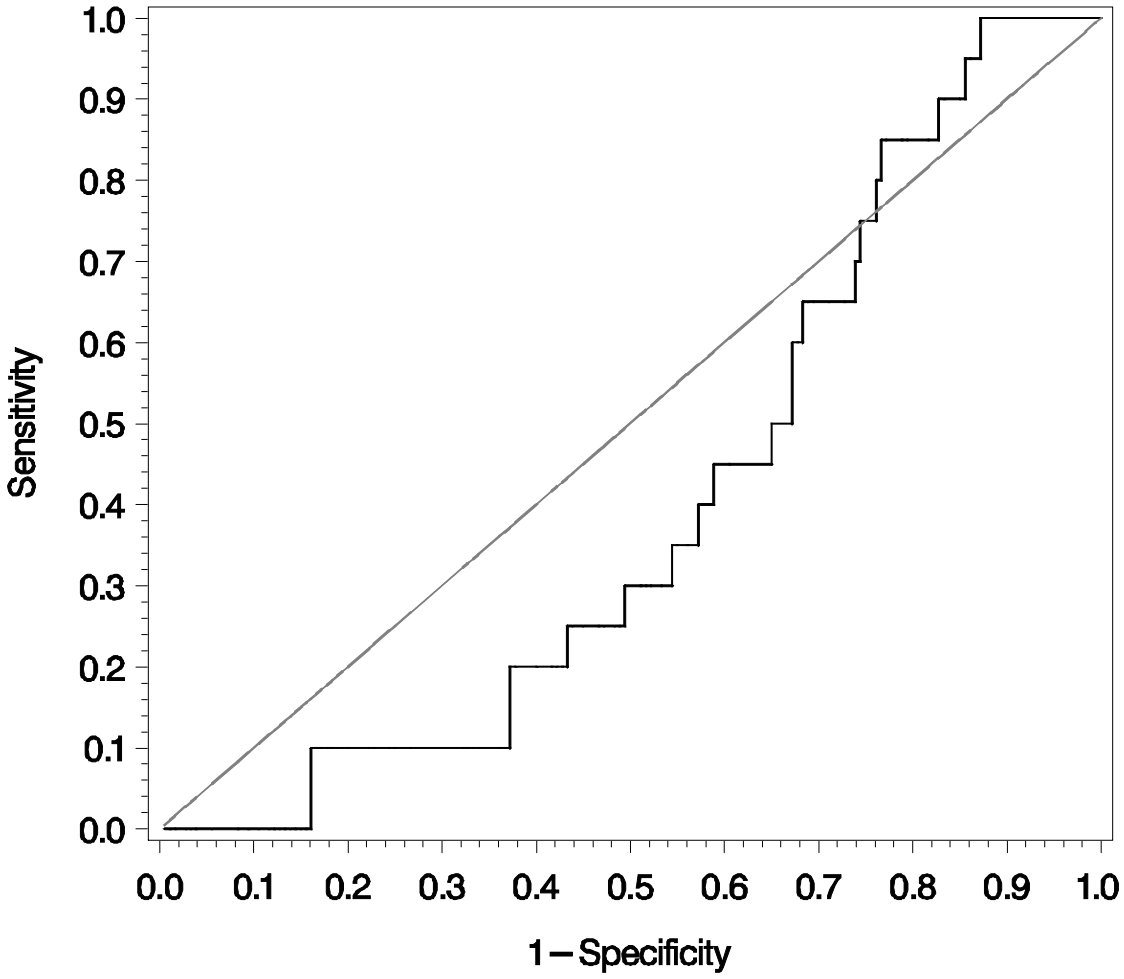


Figure 33. Mean, 5th and 95th percentile ROC plot from the results of the regression quadratic score approach.

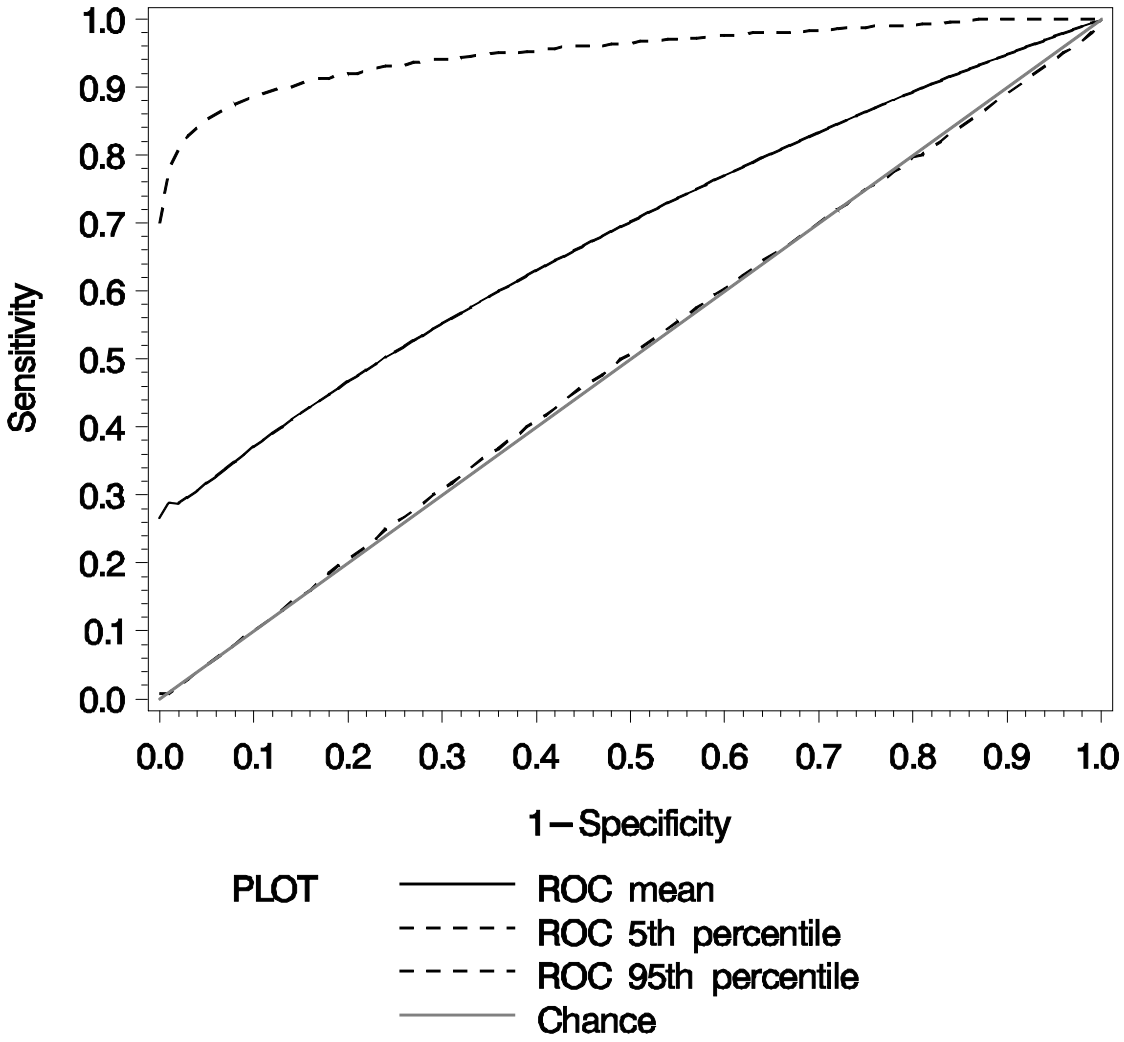


Figure 34. Mean, 5th and 95th percentile ROC plot from the results of the Bartlett's quadratic score approach.

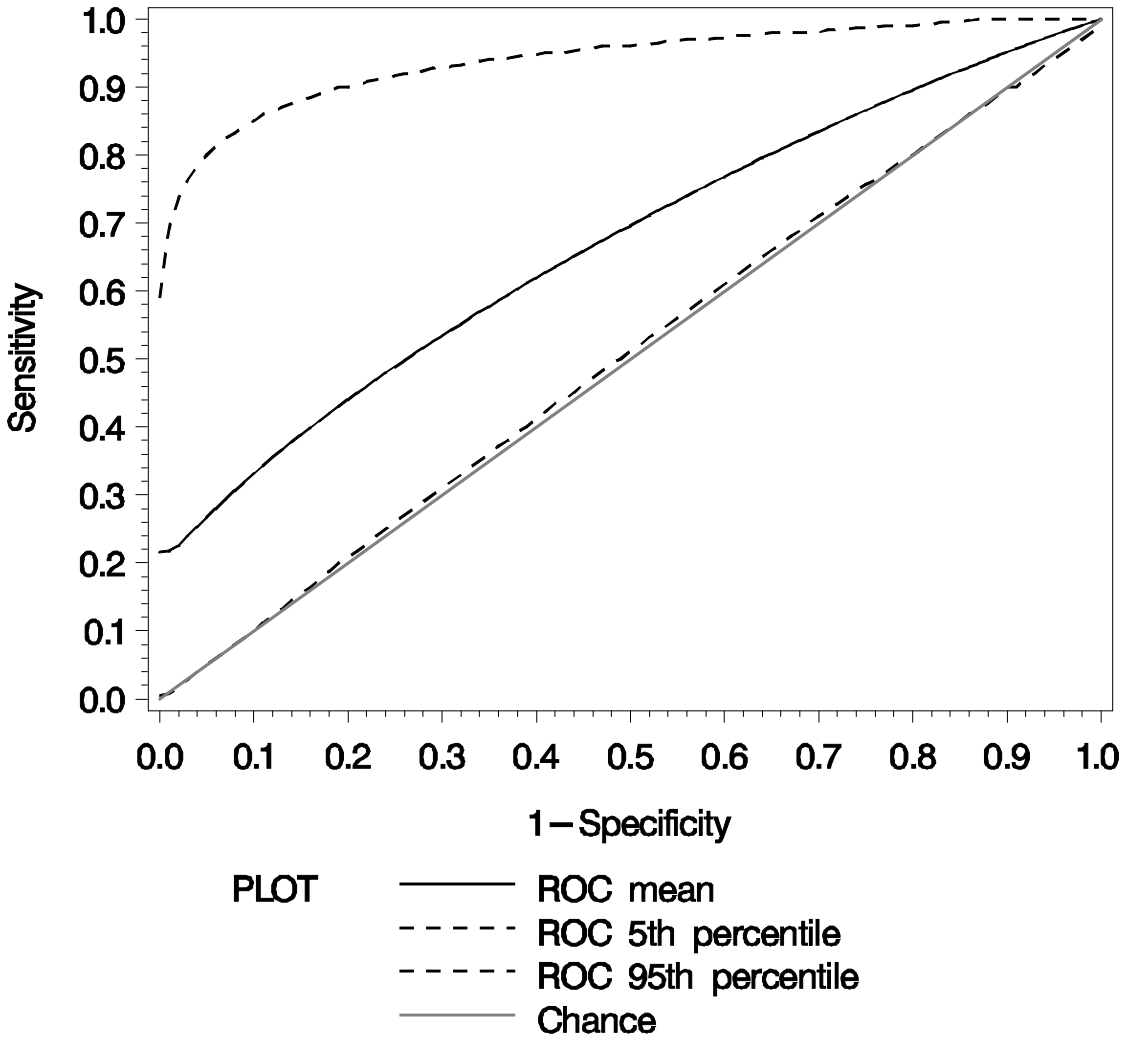


Figure 35. Mean, 5th and 95th percentile ROC plot from the results of the difference in the regression factor score residuals approach.

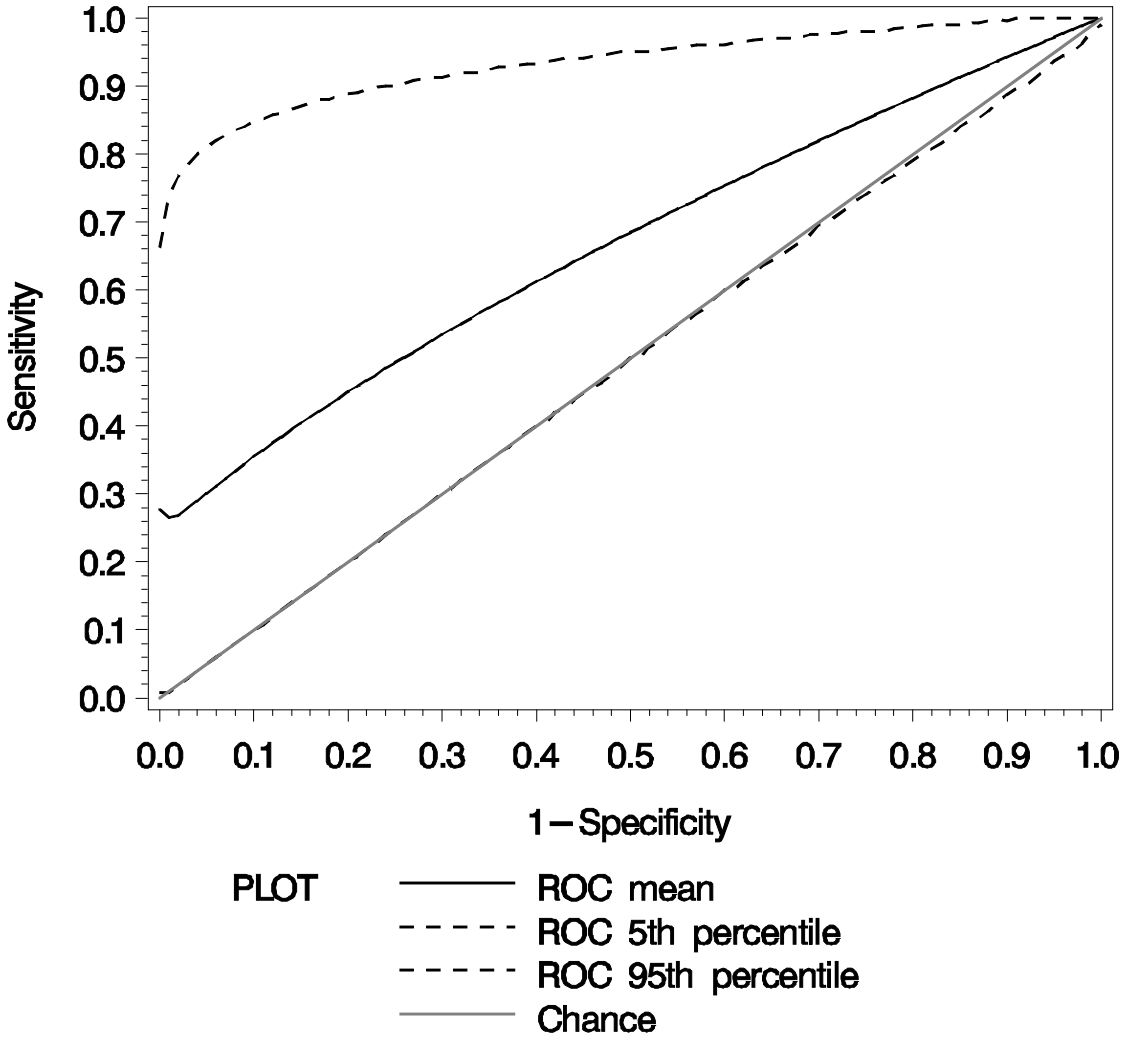


Figure 36. Mean, 5th and 95th percentile ROC plot from the results of the difference in the Bartlett's factor score residuals approach.

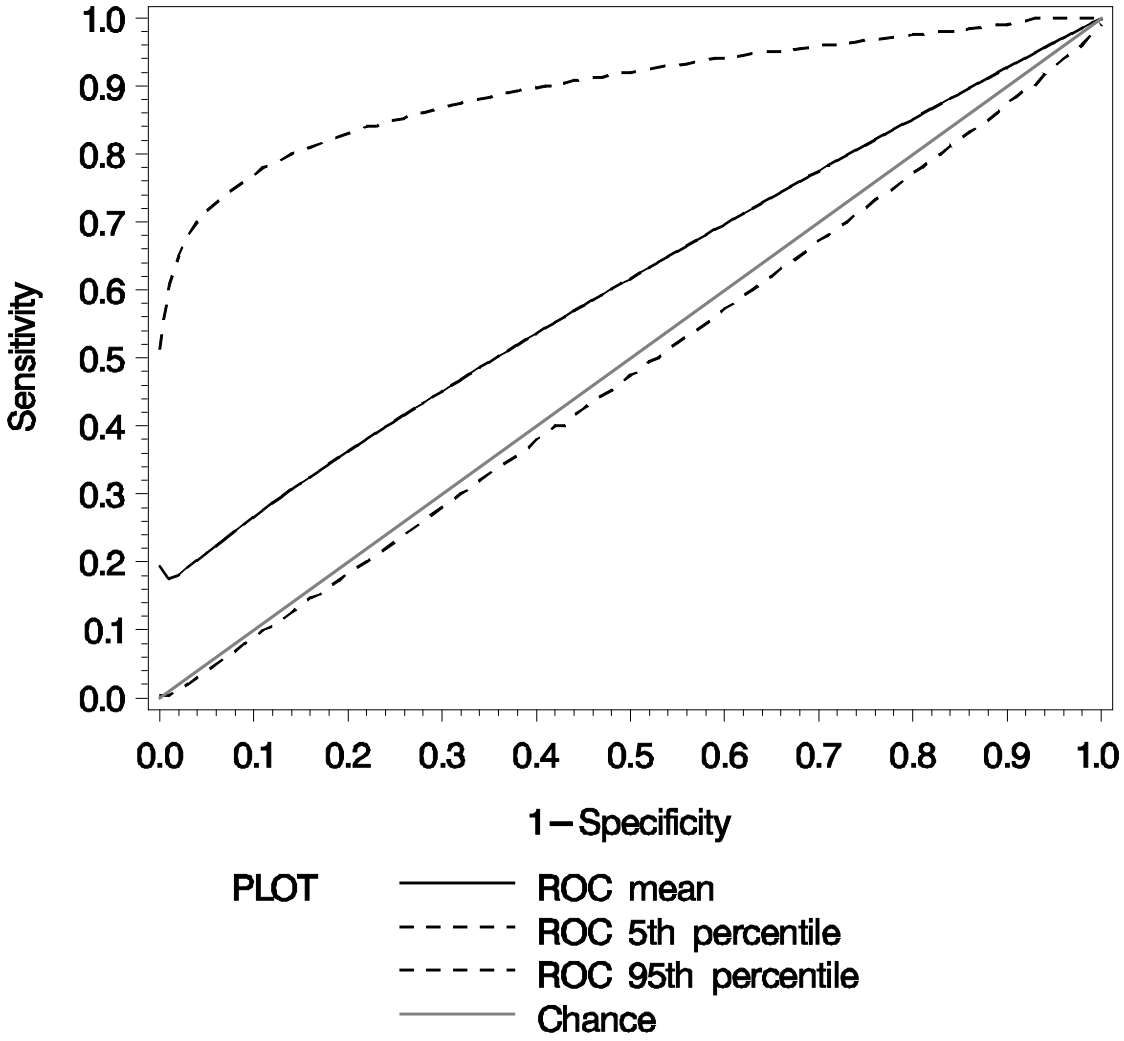


Figure 37. Mean, 5th and 95th percentile ROC plot from the results of the difference in the regression $RMSR_i$ approach.

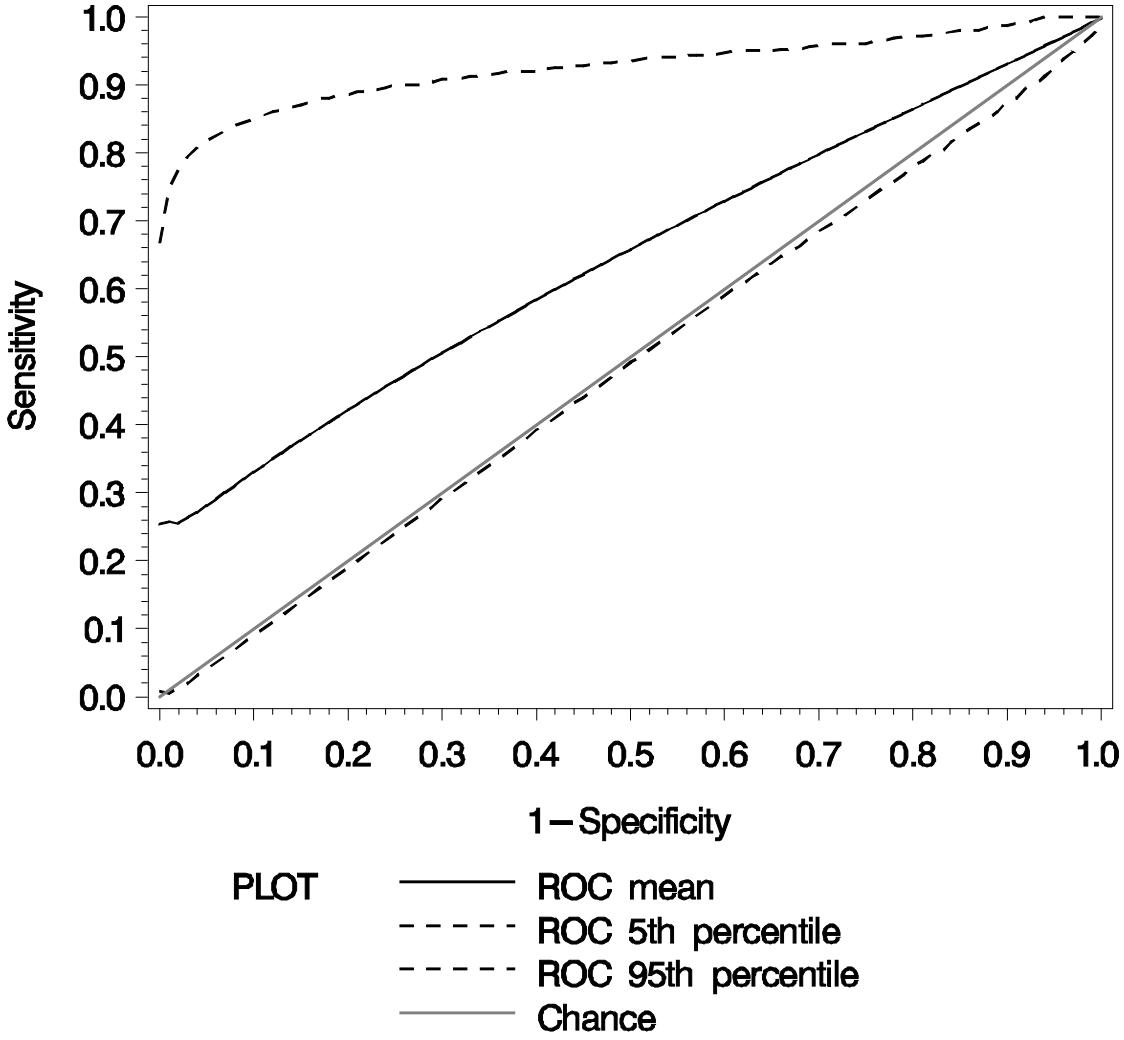


Figure 38. Mean, 5th and 95th percentile ROC plot from the results of the difference in the Bartlett's $RMSR_i$ approach.

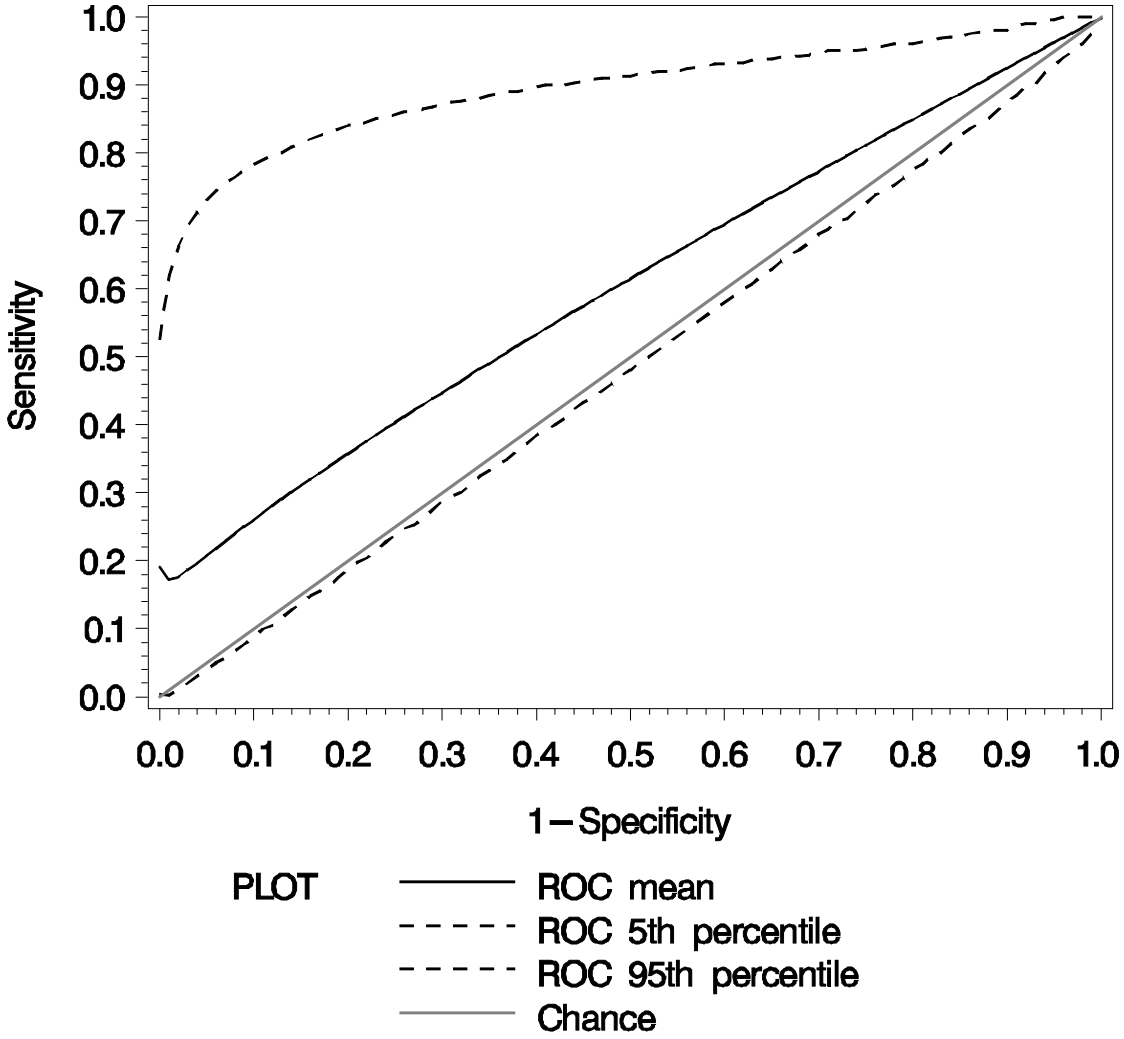


Figure 39. Mean, 5th and 95th percentile ROC plot from the results of the difference in $-2PLL_i$ approach.

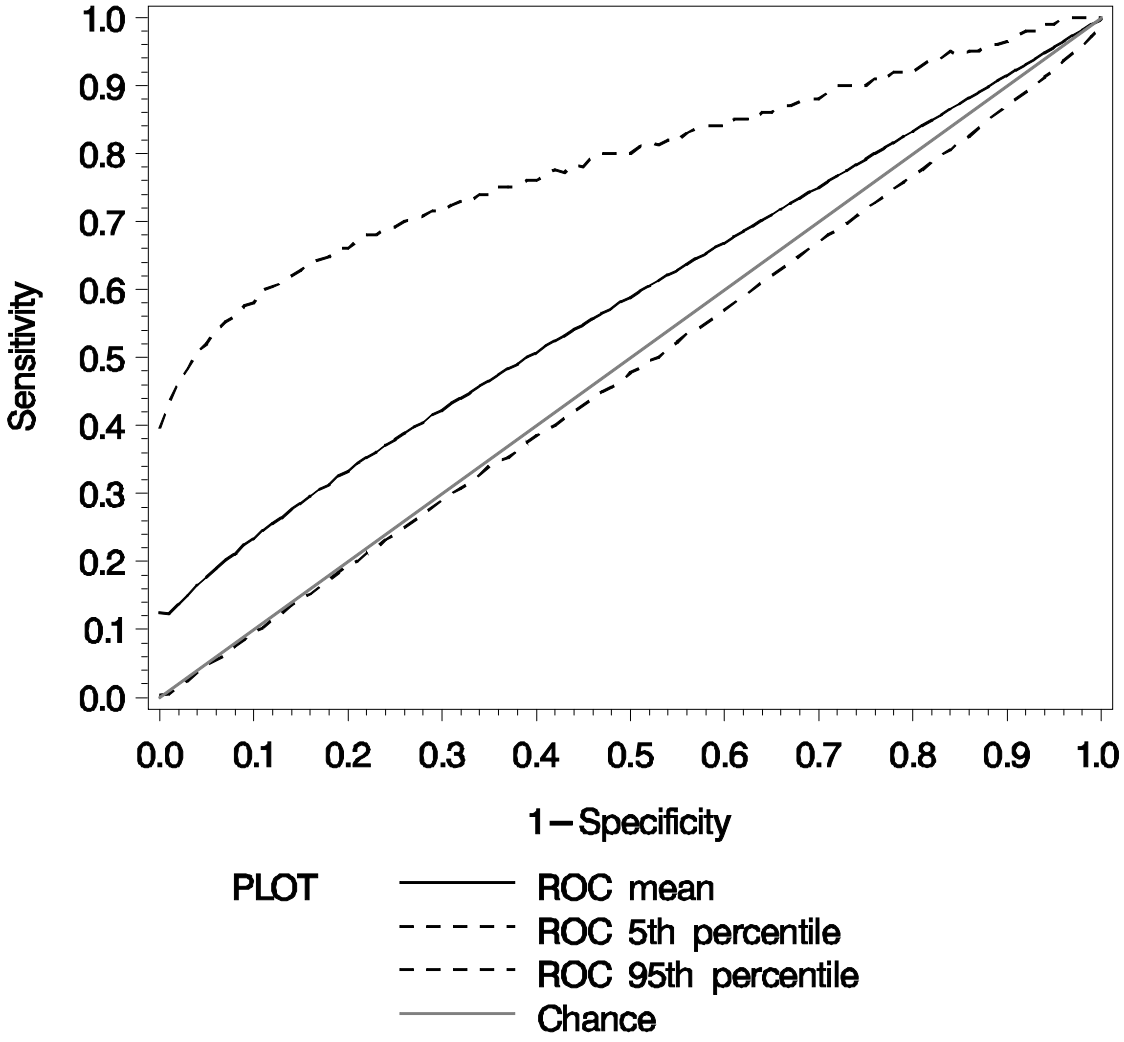


Figure 40. Mean ROC plots for each combination of the levels of number of observation times and quadratic size from the results of the regression quadratic factor score estimates approach

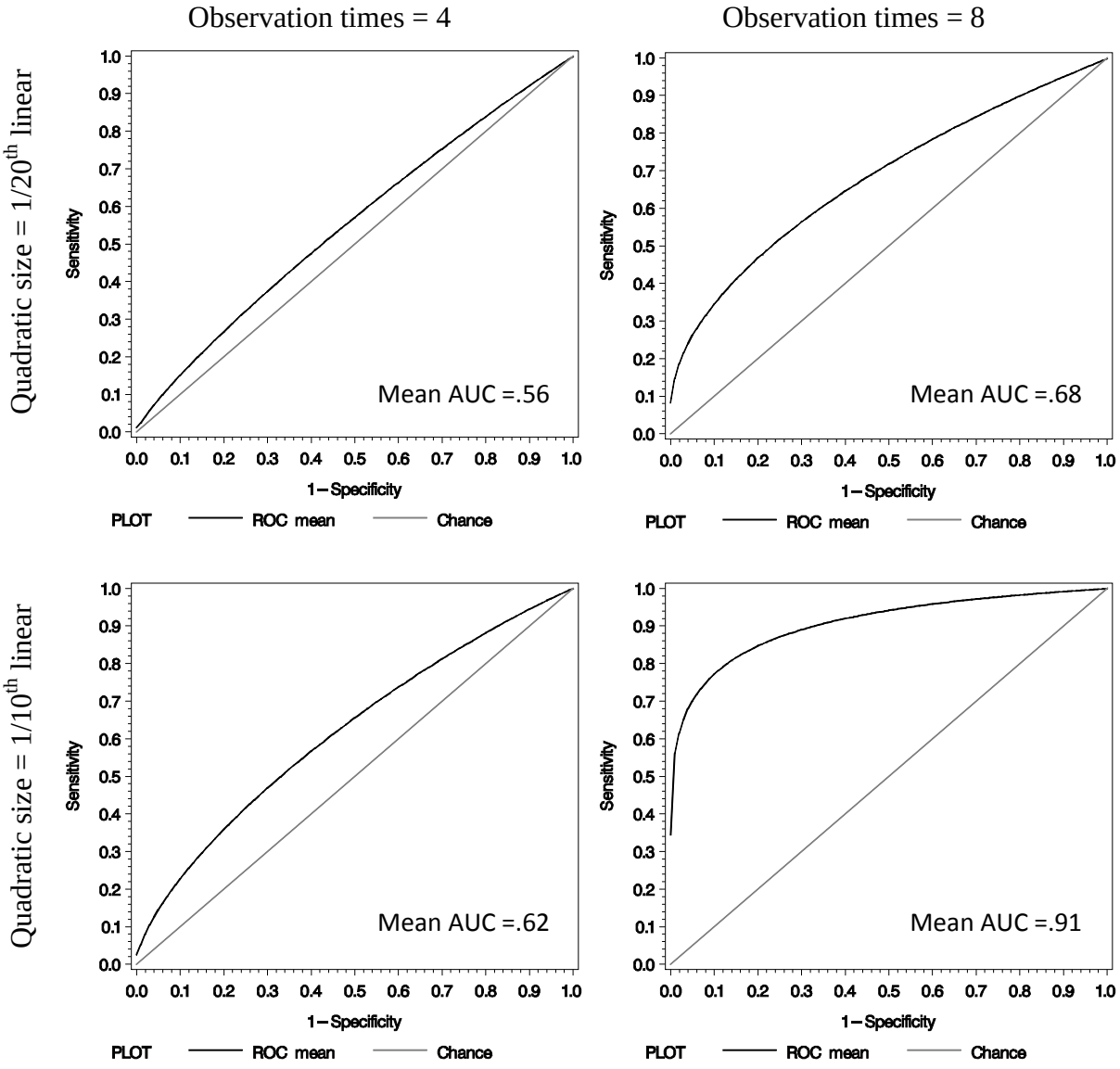


Figure 41. Mean ROC plots for each combination of the levels of number of observation times and communality from the results of the Bartlett's quadratic factor score estimates approach

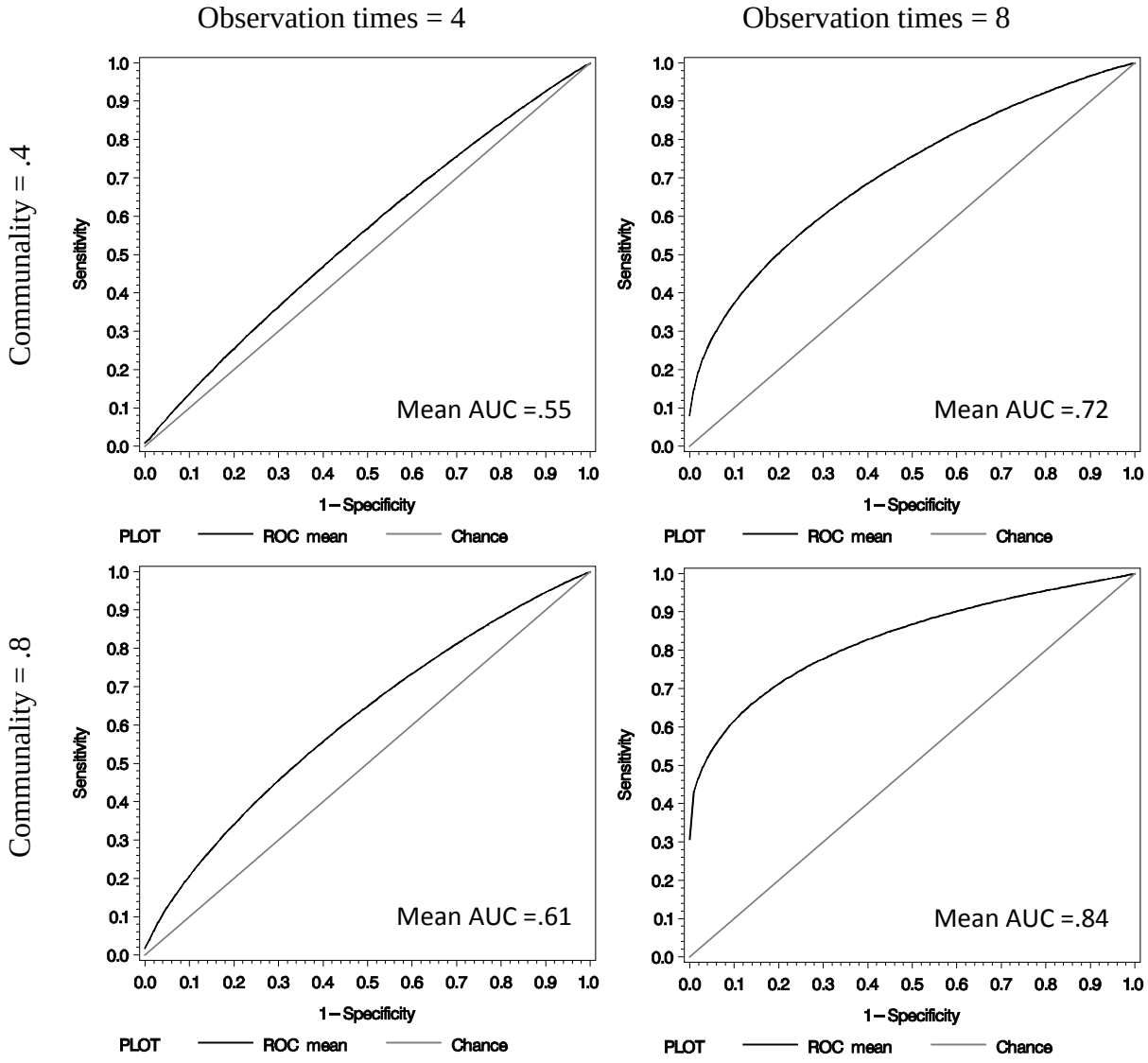


Figure 42. Mean ROC plots for each combination of the levels of number of observation times and quadratic size from the results of the Bartlett's quadratic factor score estimates approach

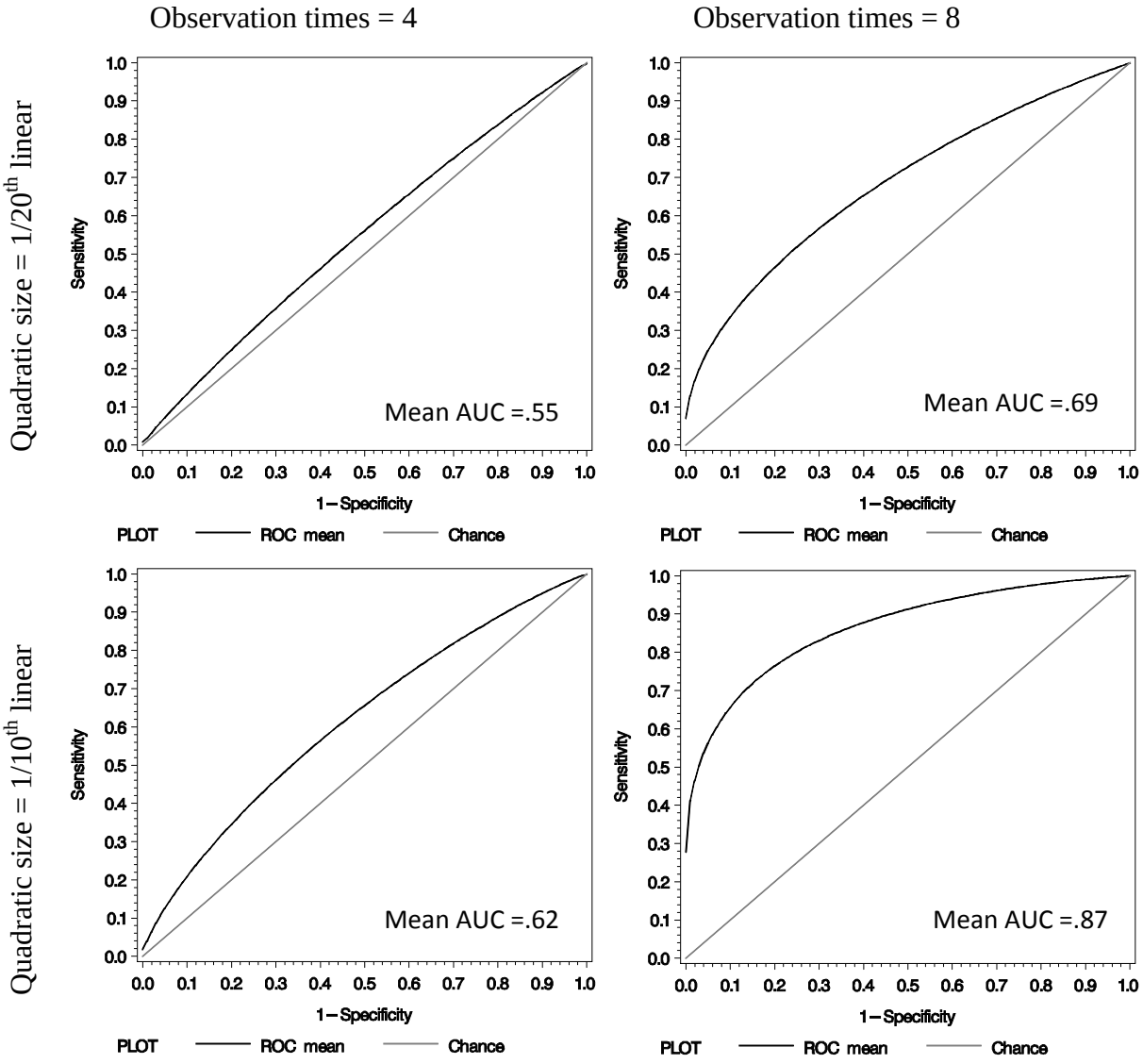


Figure 43. Mean ROC plots for each combination of the levels of number of observation times and communality from the results of the difference in regression factor score residuals approach

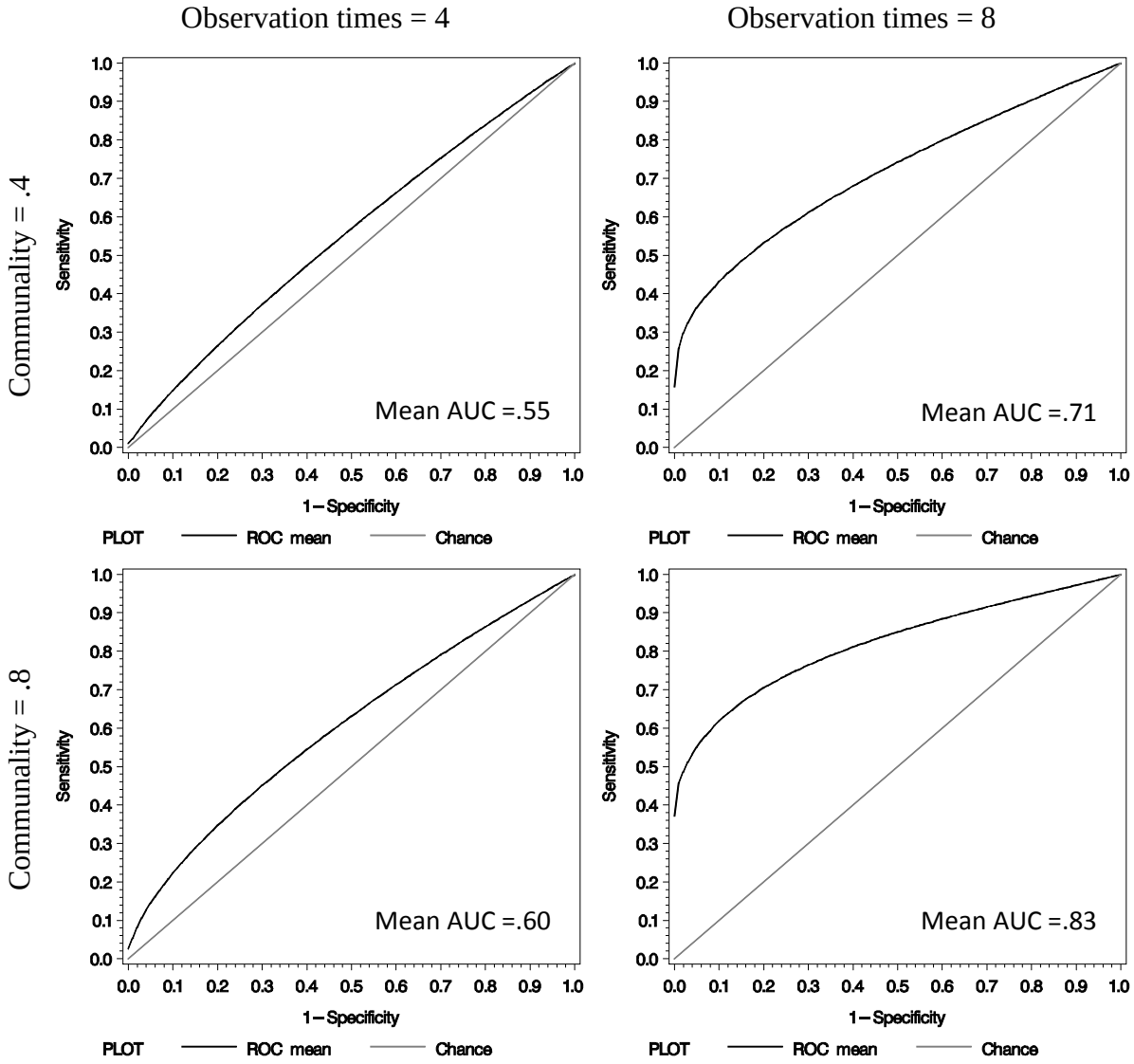


Figure 44. Mean ROC plots for each combination of the levels of number of observation times and communality from the results of the difference in Bartlett's factor score residuals approach

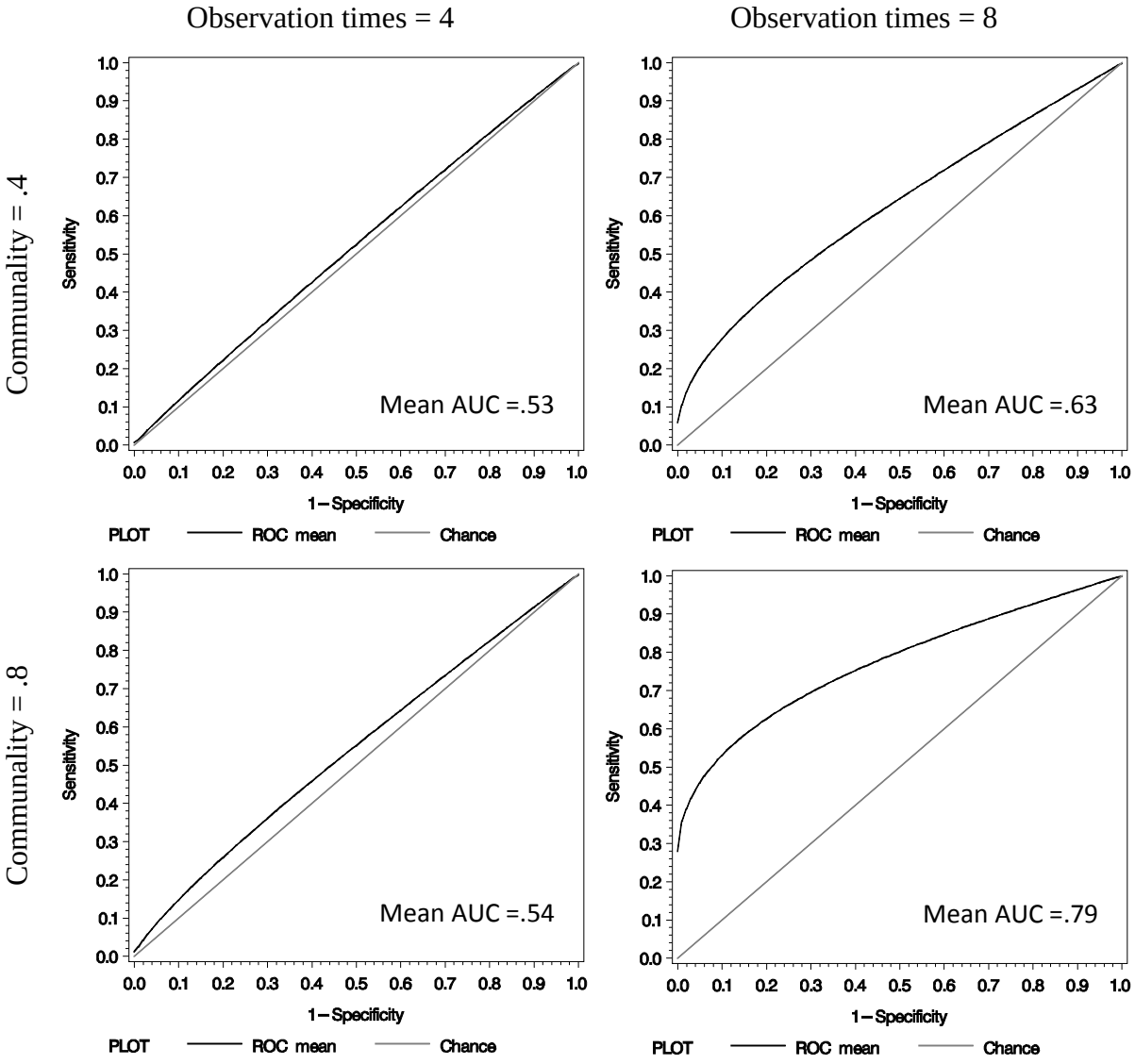


Figure 45. Mean ROC plots for each combination of the levels of number of observation times and communality from the results of the difference in the regression $RMSR_i$ approach.

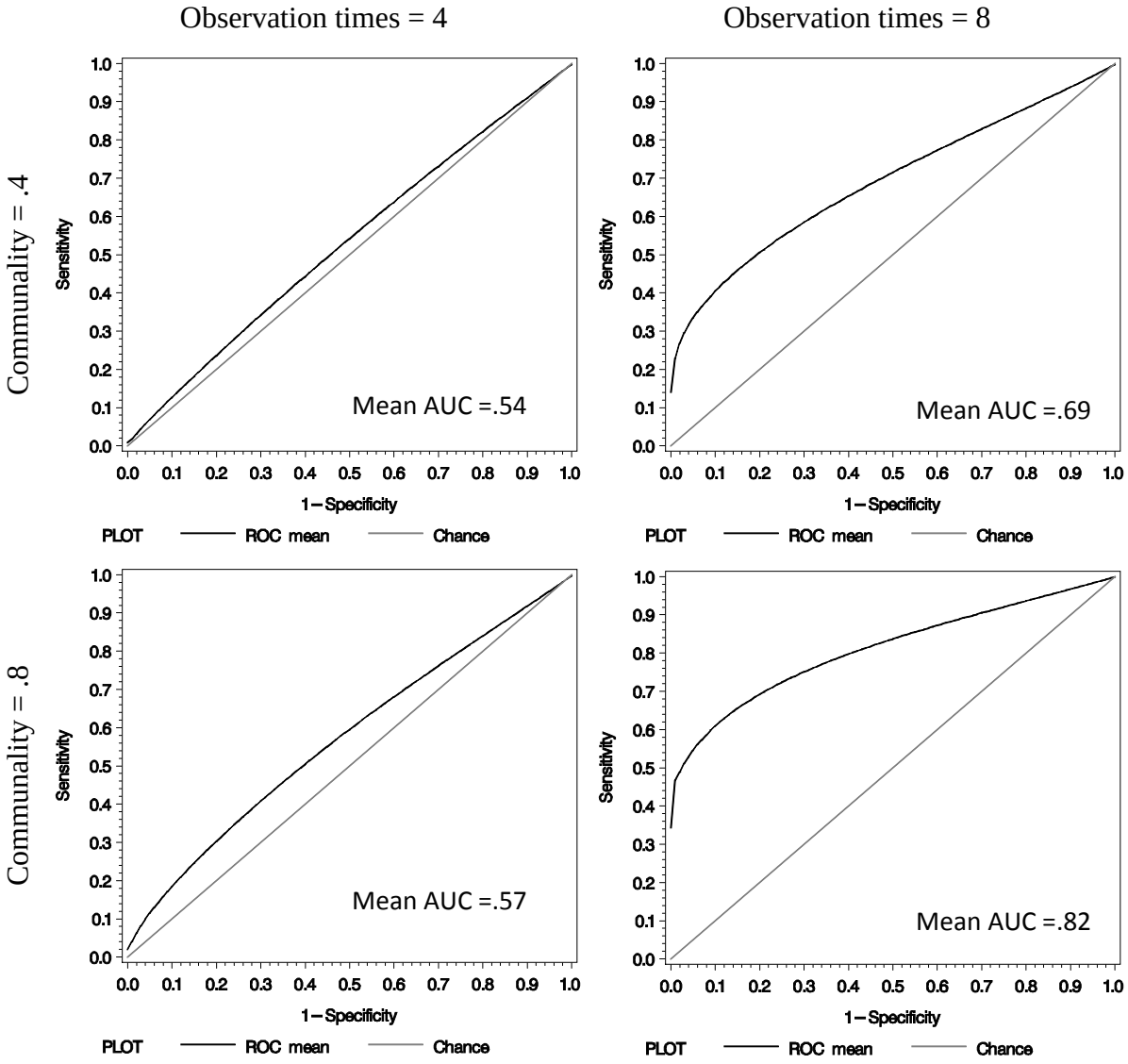


Figure 46. Mean ROC plots for each combination of the levels of number of observation times and communality from the results of the difference in the Bartlett's $RMSR_i$ approach.

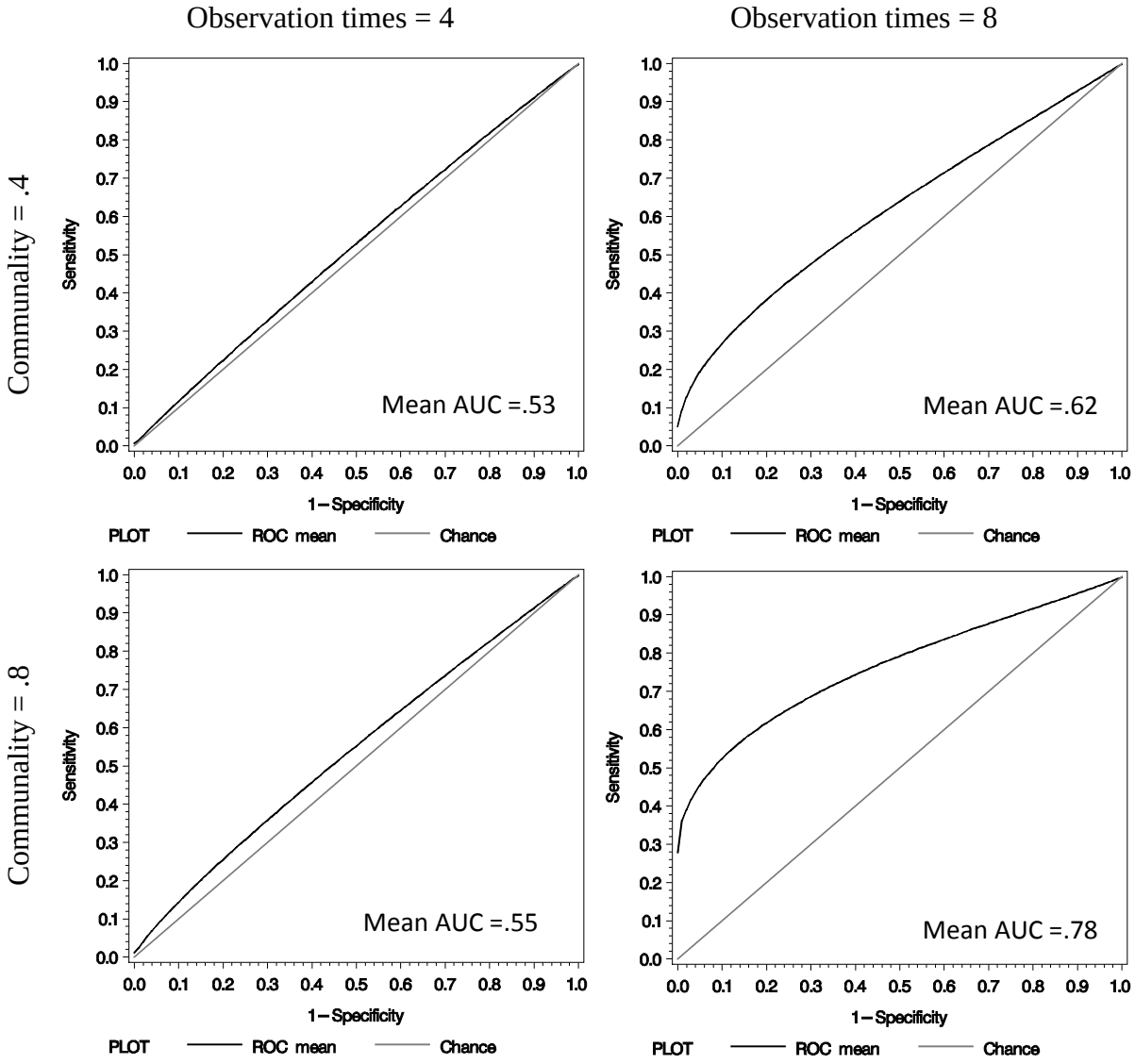


Figure 47. Mean ROC plots for each combination of the levels of number of observation times and quadratic size from the results of the difference in regression factor score residuals approach

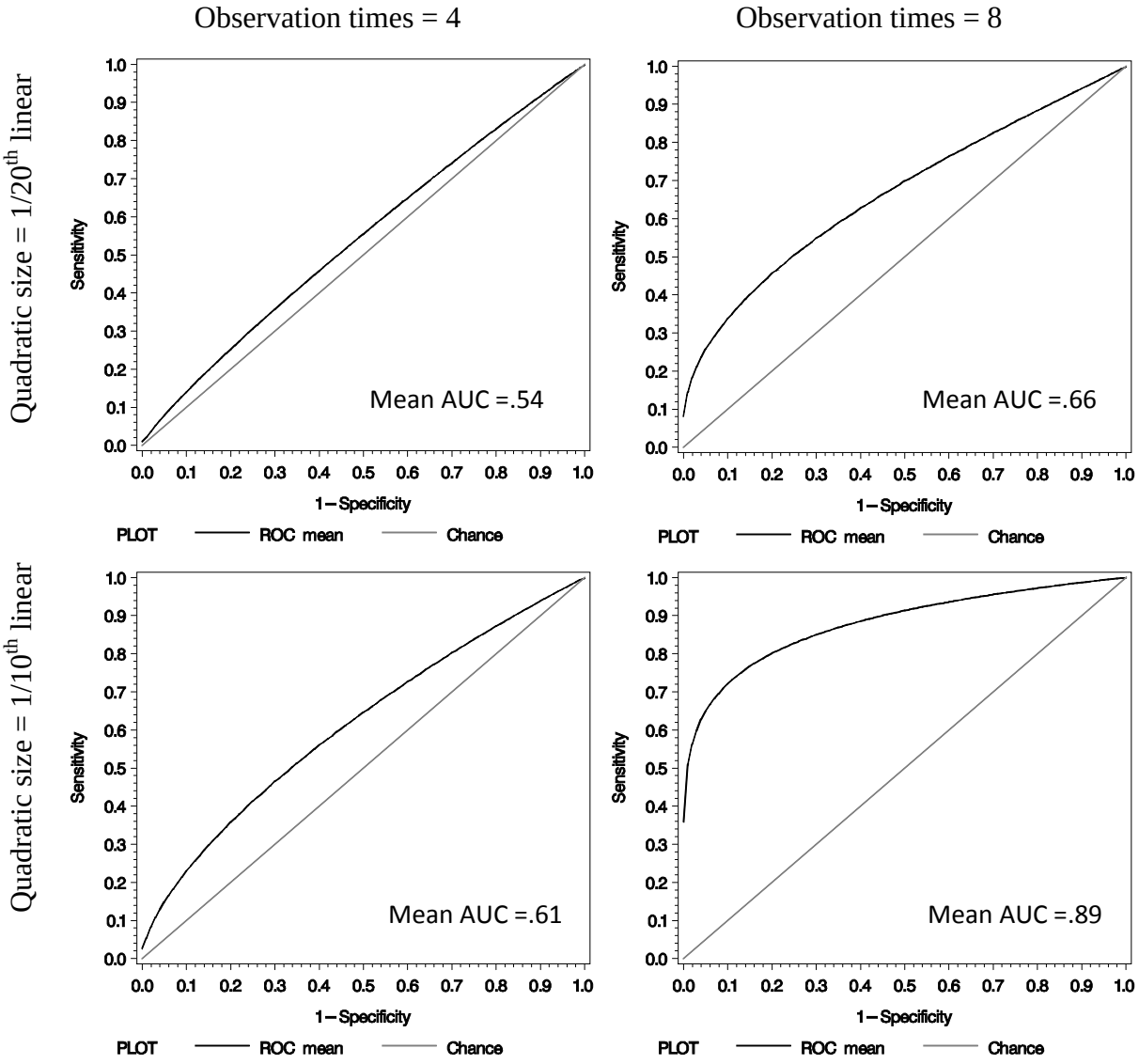


Figure 48. Mean ROC plots for each combination of the levels of number of observation times and quadratic size from the results of the difference in Bartlett's factor score residuals approach

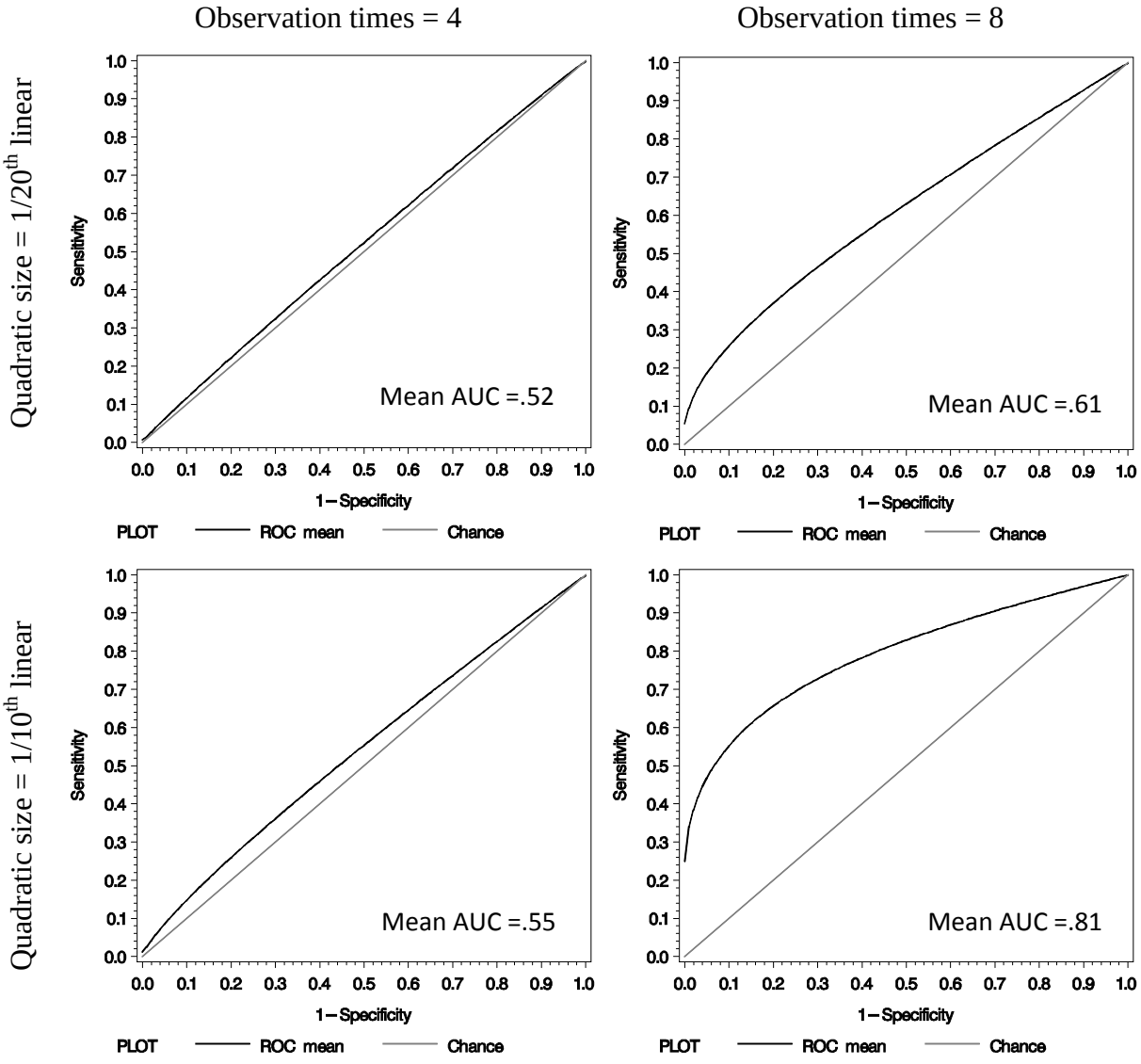


Figure 49. Mean ROC plots for each combination of the levels of number of observation times and quadratic size from the results of the difference in the regression RMSR_i approach.

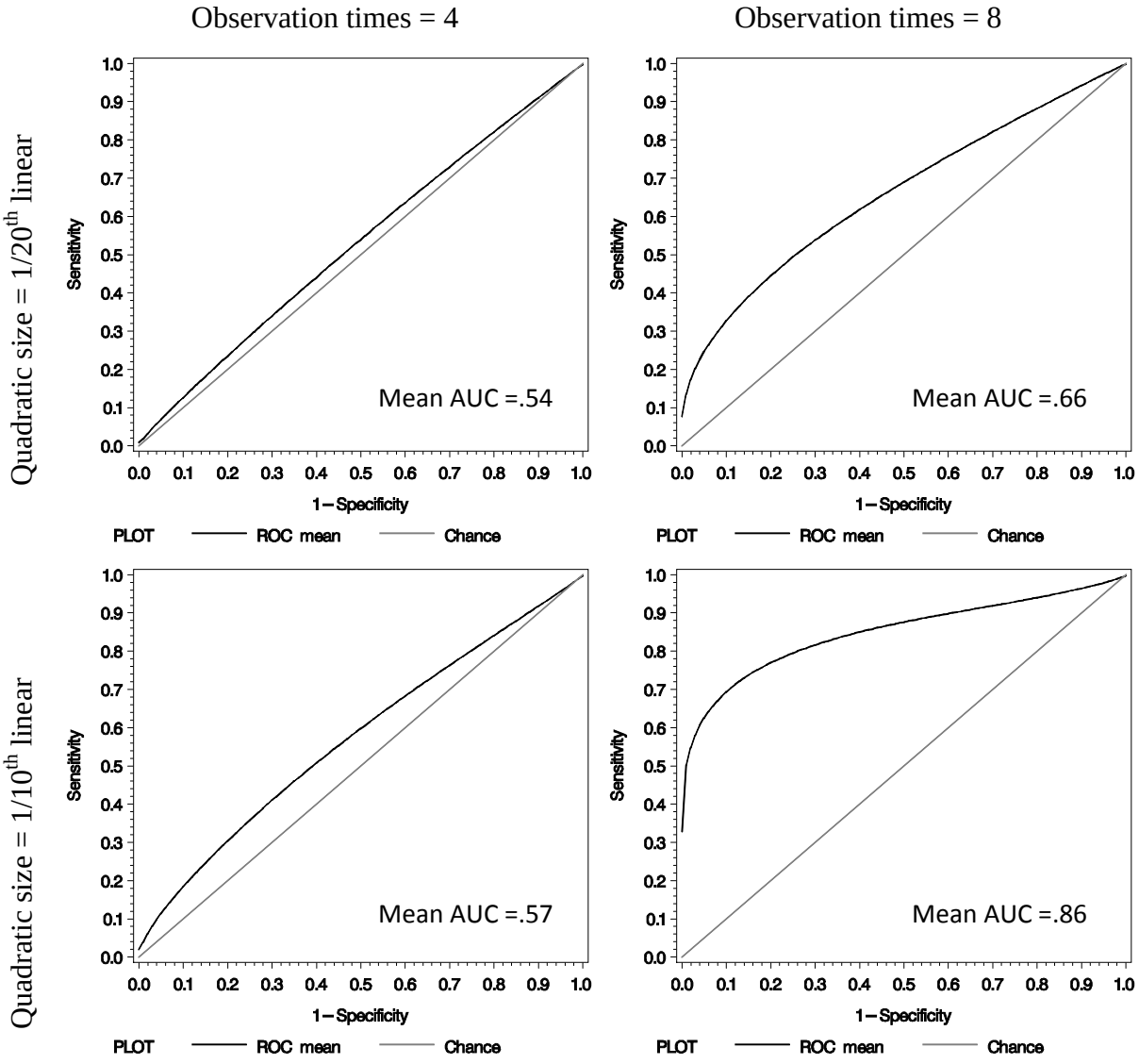


Figure 50. Mean ROC plots for each combination of the levels of number of observation times and quadratic size from the results of the difference in the Bartlett's $RMSR_i$ approach.

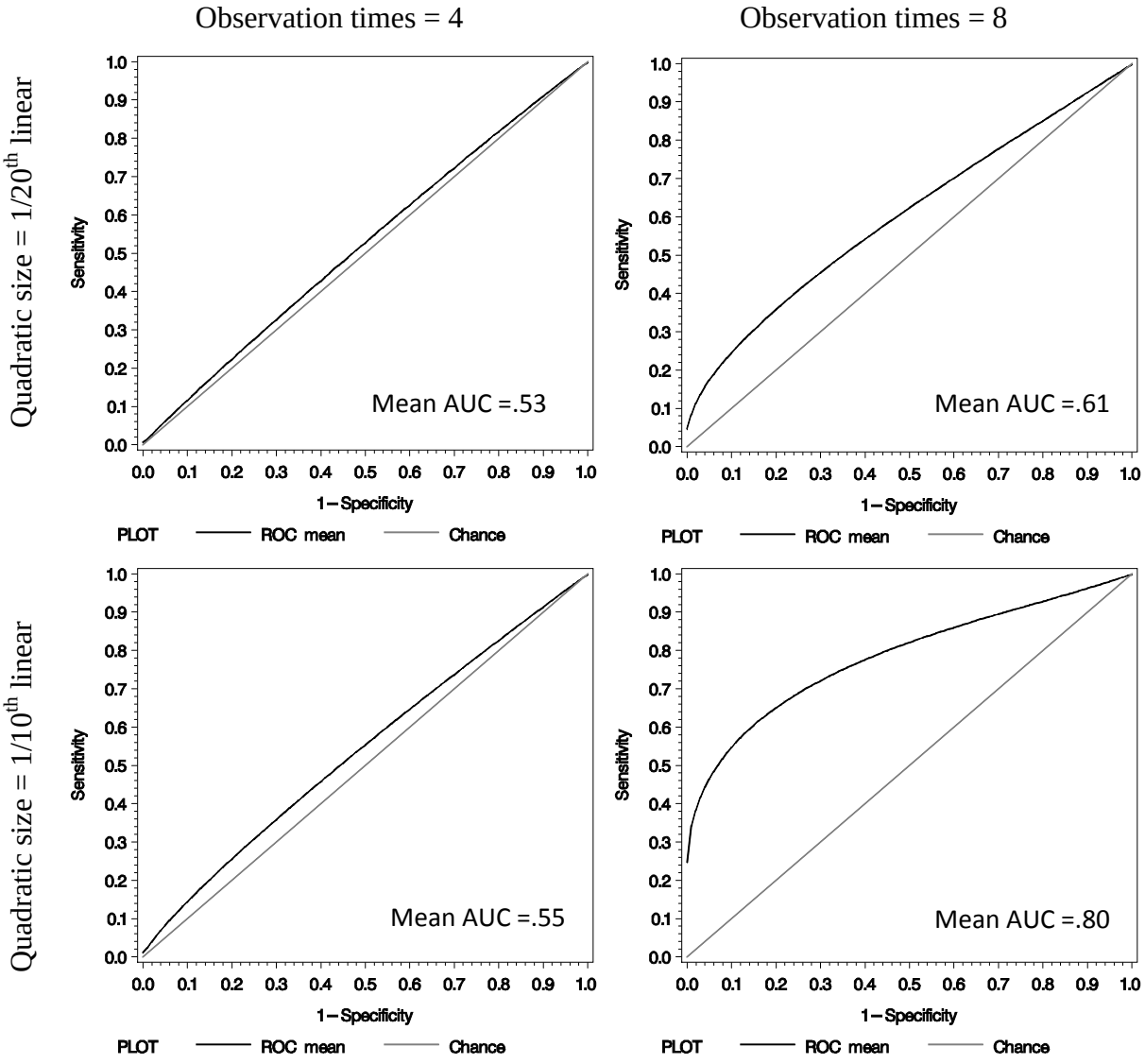


Figure 51. Mean ROC plots for each combination of the levels of communality and quadratic size from the results of the difference in the Bartlett's $RMSR_i$ approach.

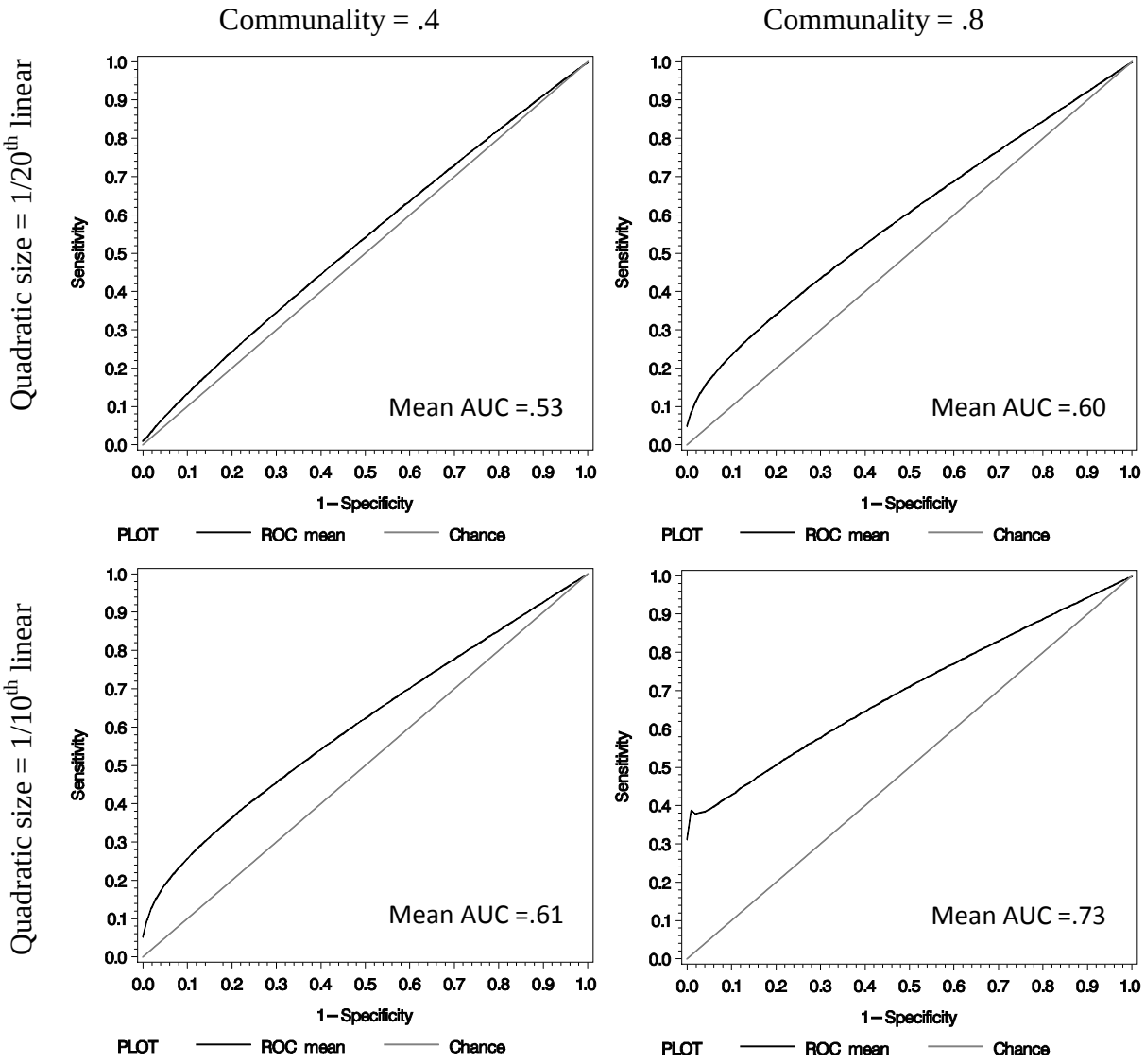


Figure 52. Mean ROC plots for each combination of the levels of number of observation times and communality from the results of the difference in -2PLL_i approach.

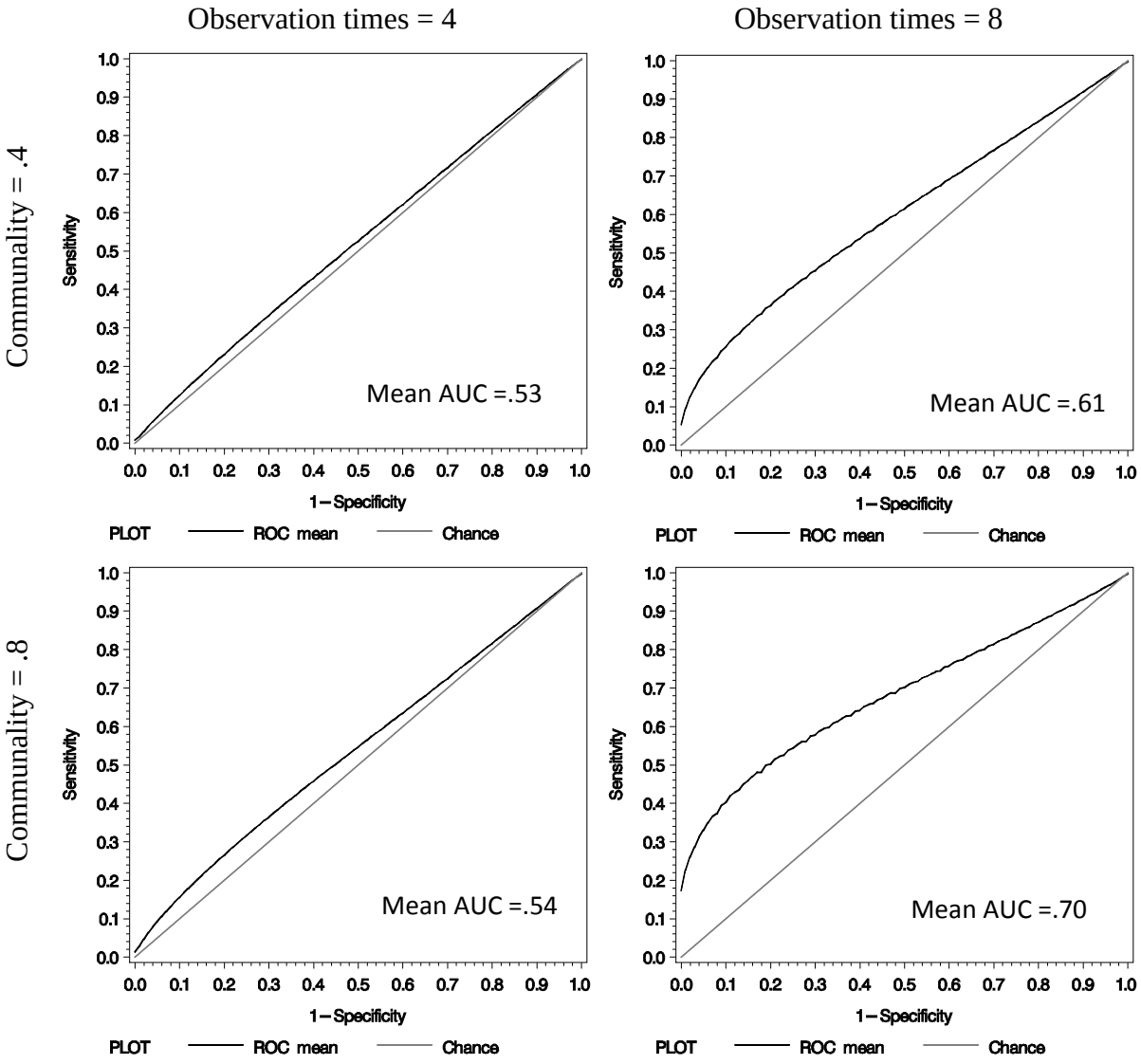


Figure 53. Mean ROC plots for each combination of the levels of number of observation times and percent of aberrant observations for the small quadratic size ($1/20^{th}$ linear) from the results of the difference in -$2PLL_i$ approach.

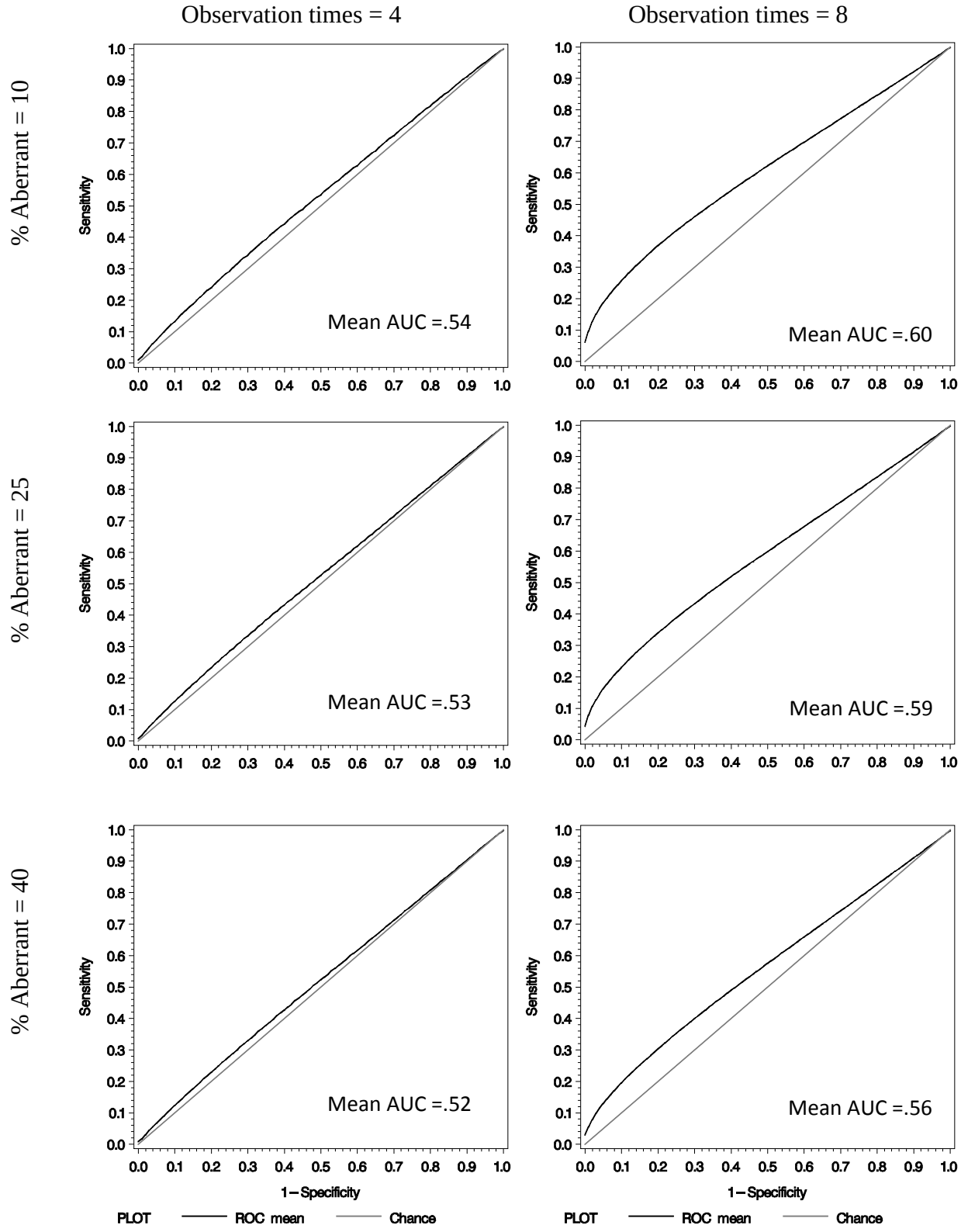
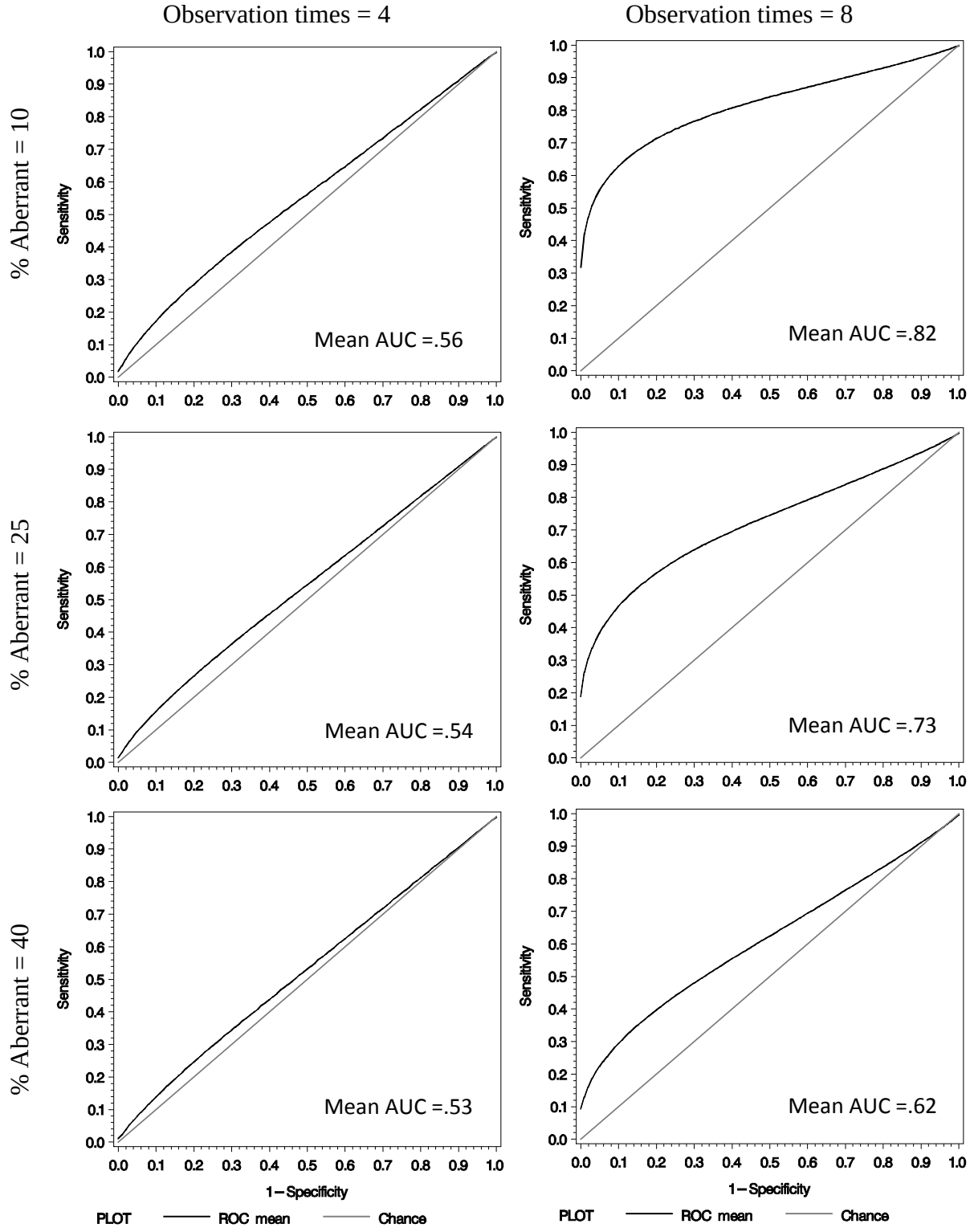


Figure 54. Mean ROC plots for each combination of the levels of number of observation times and percent of aberrant observations for the large quadratic size (1/10th linear) from the results of the difference in -2PLL_i approach.



References

Baker, P.C., Keck, C.K., Mott, F. L., & Quinlan, S. V. (1993). *NLSY Child Handbook: A guide to the 1986-1990 National Longitudinal Survey of Youth Child Data* (rev. ed.). Columbus: The Ohio State University, Center for Human Resources Research.

Bartholomew, D. J., Deary, I. J., & Lawn, M. (2009). The origin of factor scores: Spearman, Thomson and Bartlett. *British Journal Of Mathematical & Statistical Psychology*, *62*(3), 569-582. doi:10.1348/000711008X365676

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis* (2nd ed.). London: Arnold.

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology, 28,* 97-104.

Bauer, D.J. (2003). Estimating multilevel linear models as structural equation models. *Journal of Educational and Behavioral Statistics, 28*, 134-167.

Bauer, D. J., Baldasaro, R. E. & Gottfredson, N. C. (2012). Diagnostic Procedures for Detecting Nonlinear Relationships Between Latent Variables. *Structural Equation Modeling, 19*(2), 157-177. doi: 10.1080/10705511.2012.659612

Biesanz, J. C., Deeb-Sossa, N., Papadakis, A. A., Bollen, K. A., & Curran, P. J. (2004). The Role of Coding Time in Estimating and Interpreting Growth Curve Models. *Psychological Methods, 9*(1), 30-52. doi:10.1037/1082-989X.9.1.30

Blozis, S. A., & Cho, Y. (2008). Coding and Centering of Time in Latent Curve Models in the Presence of Interindividual Time Heterogeneity. *Structural Equation Modeling, 15*(3), 413-433. doi:10.1080/10705510802154299

Bollen, K. A. (1987). Outliers and improper solutions: A confirmatory factor analysis example. *Sociological Methods and Research, 15*, 375–384.

Bollen, K. A. (1989). Structural equation models with latent variables. New York, NY: Wiley.

Bollen, K. A., & Arminger, G. (1991). Observational residuals in factor analysis and structural equation models. *Sociological Methodology, 21*, 235–262.

Bollen, K.A., & Curran, P.J. (2006). Latent Curve Models: A Structural Equation Approach. Wiley Series on Probability and Mathematical Statistics.

Bollen, K. A. & Long, J. S. (Eds). (1993). *Testing structural equation models.* Newbury Park, CA: Sage.

Carrig, M. M., Wirth, R. J., & Curran, P. J. (2004). A SAS Macro for Estimating and Visualizing Individual Growth Curves. *Structural Equation Modeling, 11*(1), 132-149. doi: 10.1207/S15328007SEM1101_9

Chase-Lansdale, P. L., Mott, F. M., Brooks-Gunn, J., & Phillips, D. A. (1991). Children of the National Longitudinal Study of Youth: A unique research opportunity. *Developmental Psychology, 27*(6), 919-931.

Chou, C.-P., Bentler, P. M., & Pentz, M. A.(1998). Comparisons of two statistical approaches to study growth curves: The multilevel model and the latent curve analysis. *Structural Equation Modeling, 5*(3), 247-266. doi: 10.1080/10705519809540104

Coffman, D. L., & Millsap, R. E. (2006). Evaluating latent growth curve models using individual fit statistics. *Structural Equation Modeling, 13*, 1–27.

Curran, P.J. (2003). Have multilevel models been structural equation models all along? *Multivariate Behavioral Research, 38*, 529-569.

Curran, P. J., & Wirth, R. J. (2004). Interindividual Differences in Intraindividual Variation: Balancing Internal and External Validity. *Measurement, 2*(4), 219-227.

Drasgow, F., Levine, M. V. , & Williams, E.A. (1985) Appropriateness measurement with polychotomous item response models and standardized indices. *Journal of Mathematical and Statistical Psychology, 38*, 67-86.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed). London: Arnold.

Gorsuch, R. L. (1983). *Factor analysis.* Hillsdale, NJ: L. Erlbaum Associates.

Grice, J. W. (2001a). A comparison of factor scores under conditions of factor obliquity. *Psychological Methods*, 6(1), 67-83. doi:10.1037/1082-989X.6.1.67

Grice, J. W. (2001b). Computing and evaluating factor scores. *Psychological Methods, 6*(4), 430-450. doi:10.1037/1082-989X.6.4.430

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *The British Journal of Statistical Psychology, 8* (Part II), 65-81.

Hox, J. J. (2010). *Multilevel analysis : techniques and applications* (2nd ed). New York, NY: Routledge.

Kaplan, D. (2000). *Structural equation modeling: Foundations and extensions.* Thousand Oaks, CA: Sage Publications.

Lange, K., Westlake, J. and Spence, M. A. (1976). Extensions to pedigree analysis III. Variance components by the scoring method. *Annals of Human Genetics, 39*, 485–491. doi: 10.1111/j.1469-1809.1976.tb00156.x

Langford, I.H., & Lewis, T. (1998). Outliers in Multilevel Data. *Journal of the Royal Statistical Society, Series A*, 161 (2), 121-160.

Lee, S. Y, & Wang, S. J. (1996). Sensitivity analysis of structural equation models. *Psychometrika, 61*, 93–108.

McArdle, J. J. (1998). Modeling Longitudinal Data by Latent Growth Curve Methods. In G. A. Marcoulides, (Eds.) Modern methods for business research (pp. 359-406). Mahwah, NJ: Lawrence Erlbaum

McArdle, J. J. & Bell, R. Q. (2000). An introduction to latent growth models for developmental data analysis. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples.* (pp. 69-107). Mahwah, NJ: Lawrence Erlbaum Associates

McArdle, J. J. & Hamagami (2001). Latent difference score structural equation models for linear dynamic analyses with incomplete longitudinal data. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 107-135). Washington DC: American Psychological Association.

MacCallum, R. C., Kim, C., Malarkey, W., & Kiecolt-Glaser, J. (1997). Studying Multivariate Change Using Multilevel Models and Latent Curve Models. Multivariate Behavioral Research, 32, 215-253.

McDonald, R. P. (1974). The measurement of factor indeterminacy. *Psychometrika, 39*, 203-222.

McDonald, R. P. (2011). Measuring latent quantities. Psychometrika, 76 (4), 511-536. doi: 10.1007/s11336-011-9223-7

McDonald, R.P. & Bolt, D.M. (1998) The Determinacy of Variables in Structural Equation Models. Multivariate Behavioral Research, 33 (3), 385-401. doi: 10.1207/s15327906mbr3303_4

Mehta, P. D., & West, S. G. (2000). Putting the individual back into individual growth curves. *Psychological Methods*, *5*(1), 23-43. doi:10.1037/1082-989X.5.1.23

Meredith, W., & Tisak, J. (1990). Latent curve analysis. Psychometrika, 55, 107-122.

Molenaar, P. C. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement*, 2(4), 201-218.

Mulaik, S. A. (2009) Linear causal modeling with structural equations. Boca Raton, FL: CRC Press

Muthen, B.O. (2001). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. In Collins, L.M. & Sayer, A. (Eds.), New Methods for the Analysis of Change (pp. 291-322). Washington, D.C.: APA.

Muthen, B. O. (2002). Beyond SEM: General latent variable modeling. Behaviormetrika, 29, 81-117.

Muthen, B. O., & Muthen, L. K. (2000). Integrating person-centered and variable-centered analyses: Growth Mixture Modeling with Latent Trajectory Classes. Alcoholism: Clinical and Experimental Research, 24, 882-891.

Nagin, D. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. Psychological Methods, 4, 139-157.

Nagin, D. S. (2005). Group-based modeling of development. Cambridge, MA Harvard University Press.

Nunnally, J. C. (1978). Psychometric Theory. New York, NY: McGraw-Hill.

Preacher, Wichman, MacCallum & Briggs (2008) Latent Growth Curve Modeling. SAGE book.

Pek, J., & MacCallum, R. C. (2011). Sensitivity Analysis in Structural Equation Models: Cases and Their Influence. Multivariate Behavioral Research, 46 (2), 202-228. DOI: 10.1080/00273171.2011.561068

Raudenbush, S. W., & Bryk, A. S. (2002) *Hierarchical Linear Models : Applications and Data Analysis Methods.* Thousand Oaks: Sage Publications.

Raykov, T. & Penev, S. (2002) In G. A. Marcoulides, & I. Moustaki, (Eds.) Latent variable and latent structure models. Mahwah, NJ: Lawrence Erlbaum Associates.

Reise, S. P., & Widaman, K. F. (1999). Assessing the fit of measurement models at the individual level: A comparison of item response theory and covariance structure approaches. Psychological Methods, 4, 3–21.

SAS institute INC. (2008). *SAS/IML 9.2 User's Guide.* Cary, NC: SAS Institute Inc.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling.* Mahwah, NJ: Lawrence Erlbaum Associates.

Sideris, J. (2006). *A likelihood-based approach to detecting aberrant individuals in confirmatory factor analytic models* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3219129)

Skrondal, A., & Rabe-Hesketh, S., (2004). *Generalized latent variable modeling: multilevel, longitudinal, and structural equation models.* Boca Raton: Chapman & Hall/CRC.

Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling.* Thousand Oaks, CA: Sage.

Tanaka, Y., Watadani, S., & Moon, S. H. (1991). Influence in covariance structure analysis: With an application to confirmatory factor analysis. Communications in Statistics, Series A, 20, 3805–3821.

Thomson, G. H. (1936). Some points of mathematical technique in the factorial analysis of ability. *Journal of Educational Psychology, 27,* 37-54.

Thomson, G. H. (1951). *The factorial analysis of human ability* (5th edition). Boston: Houghton Mifflin.

Thurstone, L. L. (1935). *Vectors of the mind.* Chicago, IL: University of Chicago Press.

Tomarken, A. J., & Waller, N. G. (2005). Structural Equation Modeling: Strengths, Limitations, and Misconceptions. Annual Review of Clinical Psychology, 1, 31-65. doi: 10.1146/annurev.clinpsy.1.102803.144239

Tucker (1971) Relations of factor score estimates to their use. Psychometrika, 36, 427-436. doi: 10.1007/BF02291367

Verbeke, G. & Molenburghs, G. (2000). *Linear mixed models for longitudinal data.* New York: Springer.

Ke-Hai, Y., Wing Kam, F., & Reise, S. P. (2004). Three Mahalanobis distances and their role in assessing unidimensionality. *British Journal Of Mathematical & Statistical Psychology, 57*(1), 151-165.

Yung, Y. (1997). Finite mixtures in confirmatory factor-analysis models. *Psychometrika, 62*, 297-330.